CS5604: Information Storage and Retrieval

Collection Management Tobacco Settlement Documents

December 5, 2019

Virginia Tech, Blacksburg, VA - 24061

Instructor: Dr. Edward A. Fox

TA: Ziqian Song

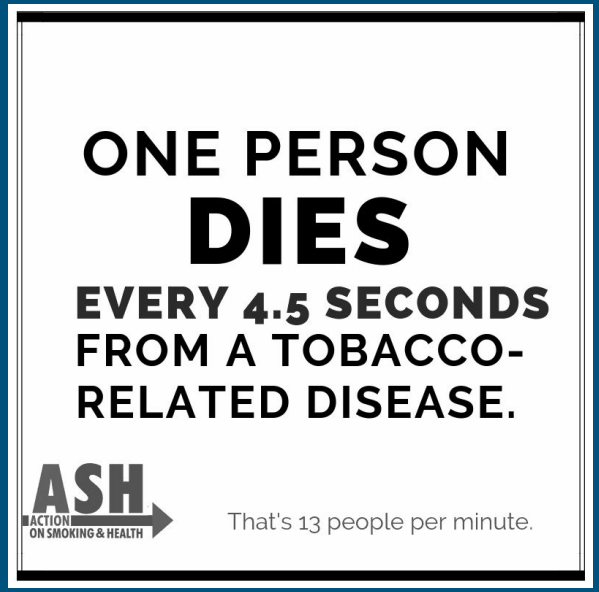
Team: Alon Bendelac, Andrei Svetovidov, Ashin Marin Thomas, Debasmita Biswas, Sushmethaa Muhundan, Yan Zhao

Outline

- Background
- Objective
- Approach
- Metadata Processing
- Text Processing
- OCR Comparison
- Future Work
- References

Background

How Do You Sell Death?



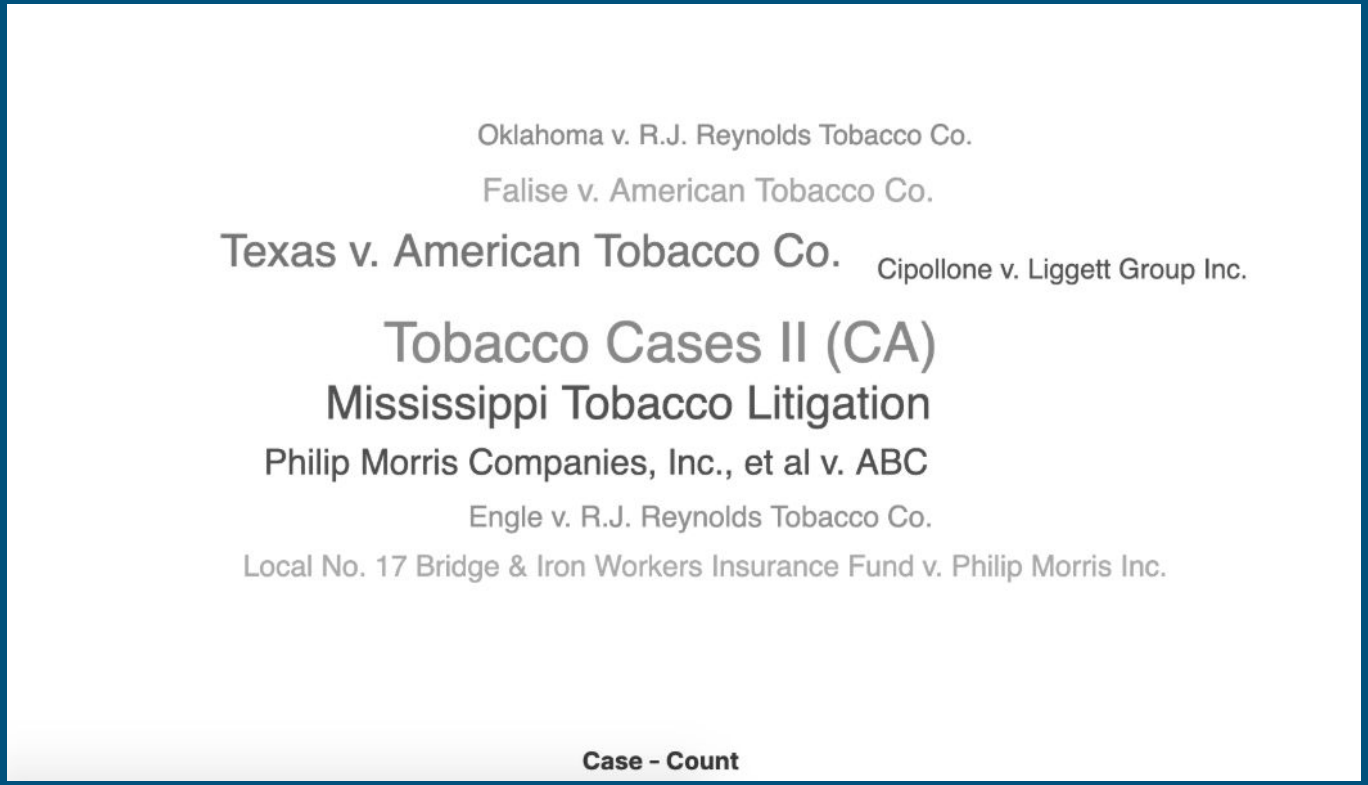
Dr. Townsend's Research

- Understand the organizational strategies and corporate tactics employed by tobacco companies to fight through the Tobacco Settlement Cases
- Be able to construct a timeline and uncover the most prominent characters and analyze the roles they have played in these cases

UCSF Industry Documents Library

- Five industries: tobacco, drug, chemical, food, fossil fuel
- 14M tobacco documents that were produced during litigation between 39 states and the seven major tobacco industry organizations
- Bulk of documents date between 1950 - 2010
- Documents have been OCR’ed and managed using MySQL database

Top 10 cases


Oklahoma v. R.J. Reynolds Tobacco Co.
Falise v. American Tobacco Co.
Texas v. American Tobacco Co.
Cipollone v. Liggett Group Inc.
Tobacco Cases II (CA)
Mississippi Tobacco Litigation
Philip Morris Companies, Inc., et al v. ABC
Engle v. R.J. Reynolds Tobacco Co.
Local No. 17 Bridge & Iron Workers Insurance Fund v. Philip Morris Inc.
Case - Count

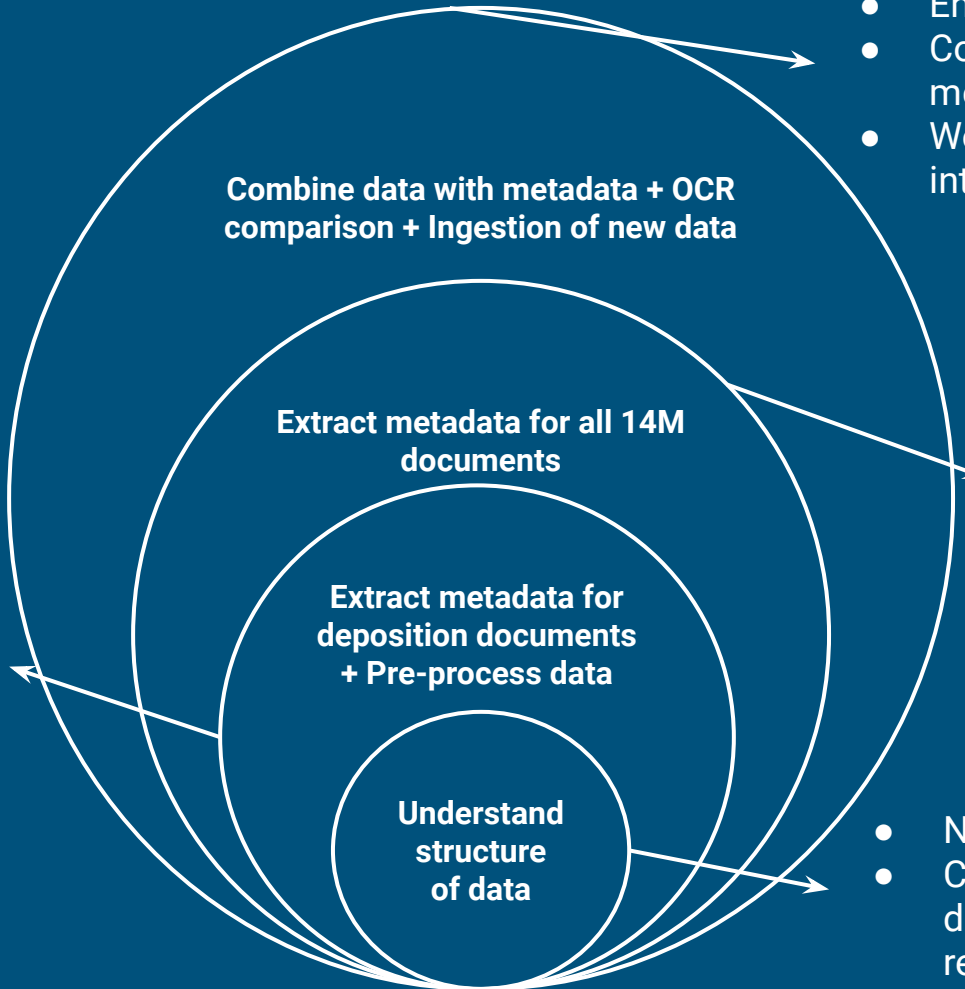
Objective

- Process the 14 million Tobacco Settlement Documents
- Clean and extract meaningful data from the documents
- Build an efficient information retrieval system
- Aid Dr.Townsend’s research

Approach

Progress

- Determined various document types
- Created Python scripts to extract metadata for deposition documents
- Simultaneously started working on cleaning, lemmatization and tokenization of data for TML team to consume

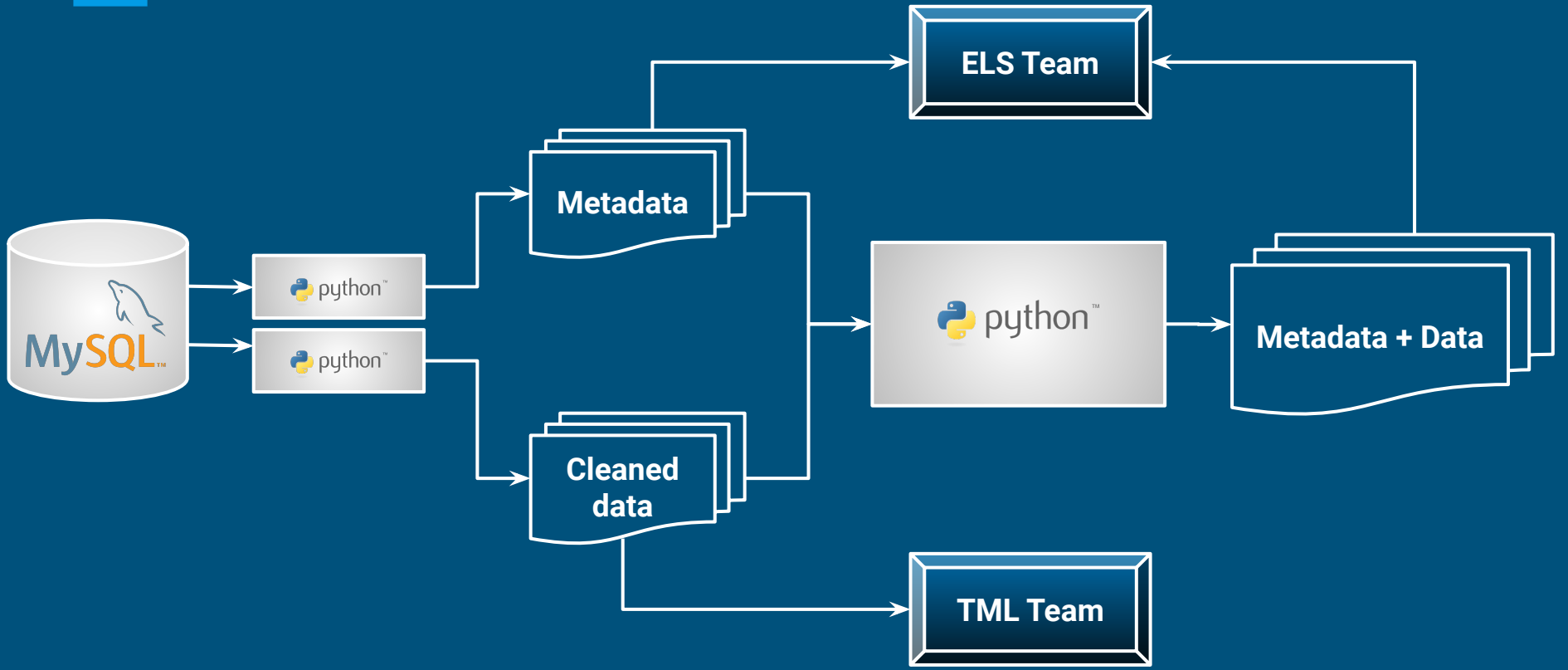


- Enable full-text searching
- Compare alternate OCR methods
- Work on ingesting new data into the system

- Close to 72,000 document types; impractical to go type by type
- Optimized our script to extract metadata 1 million at a time!

- No proper documentation
- Contacted UCSF developers for details regarding the data

Flowchart


ELS Team
Metadata
MySQL
python
python
python
Metadata + Data
Cleaned data
TML Team

Deliverables

- The metadata and data processed are stored in Ceph for the other teams to consume
- **/mnt/ceph/shared/tobacco**
  - metadata
  - data

Metadata Processing

Metadata Processing

- Overview
  - Metadata is managed in a MySQL database
  - Each document is defined by a numerical ID
  - For each document, the metadata consists of a set of key-value pairs
- Goal
  - Transform metadata from MySQL database to a collection of JSON files in Elasticsearch format

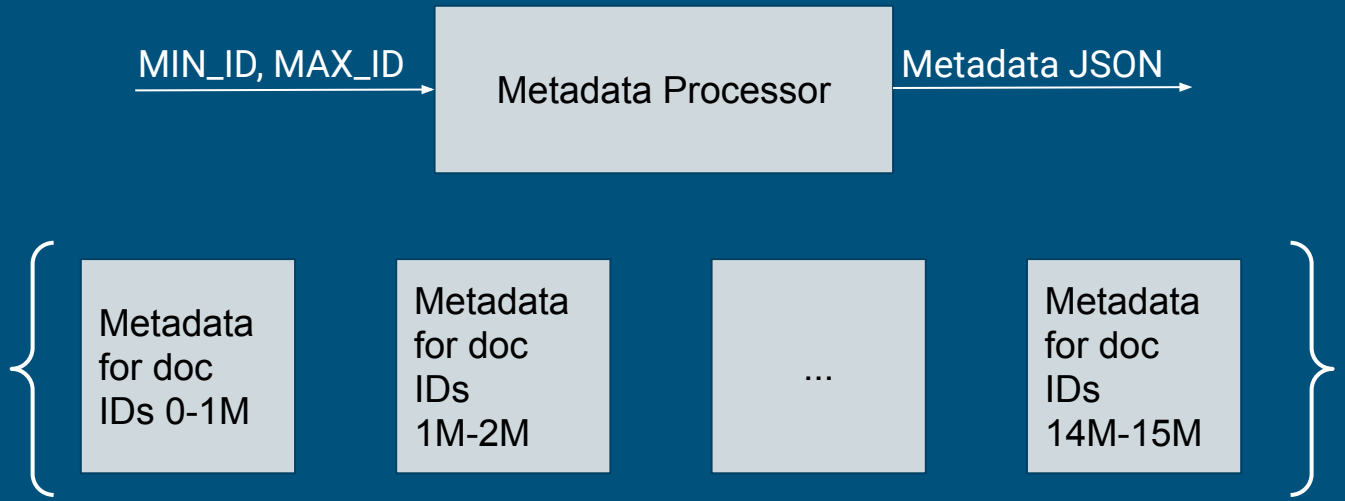
Metadata JSON Format for Elasticsearch

Header

Key-value entries


```
{"index": {"_id": 1, "_index": "tobacco"}}
{"url": "https://s3-us-west-2.amazonaws.com/edu.ucsf.library.iddl.
 artifacts/f/z/w/d/fzwd0000/fzwd0000.pdf",
 "Legacy_(LTDL2)_Tobacco_Id": "gbu00a00",
 "Title": "RISING MEDICAL COSTS REQUIRE G.H.C. DUES INCREASE IN 1967
 ↪  VIEW",
 "Document_Date": "1967-02-28 00:00:00",
 "Author": "NEWMAN HF;SIEGAL A",
 "Case": "MNAG",
 "Description": "DISCUSSES DUES INCREASE",
 "Date_Added_UCSF": "2002-02-01 00:00:00",
 "Document_Type": "article",
 "availablility": "public",
 "availablilitystatus": "no restrictions",
 "Mentioned": "GROUP HEALTH COOPERATIVE OF PUGET SOUND;...
```

Metadata Processing Approach



Metadata Processor Pseudocode

Goal: Print metadata entries for documents with ID between MIN_ID and MAX_ID

1. Connect to DB
2. Retrieve all metadata entries for documents with ID between MIN_ID and MAX_ID, ordered by ID
3. While there are entries left:
    1. Print header (identifying next document)
    2. Print “{“
    3. For each key-value pair for next doc ID, print “<key>: <value>”
    4. Print “}”

Text Processing

Text Processing - Overview

Text processing is a crucial part in preparing stored documents for further digestion by other systems. It involves:

- Text cleaning
- Text preprocessing
- Format adjustment

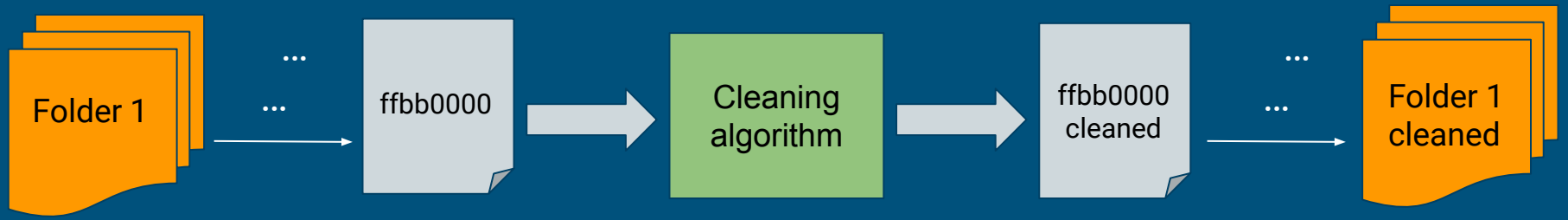
Document Text Cleaning - Overview

Why cleaning is necessary?

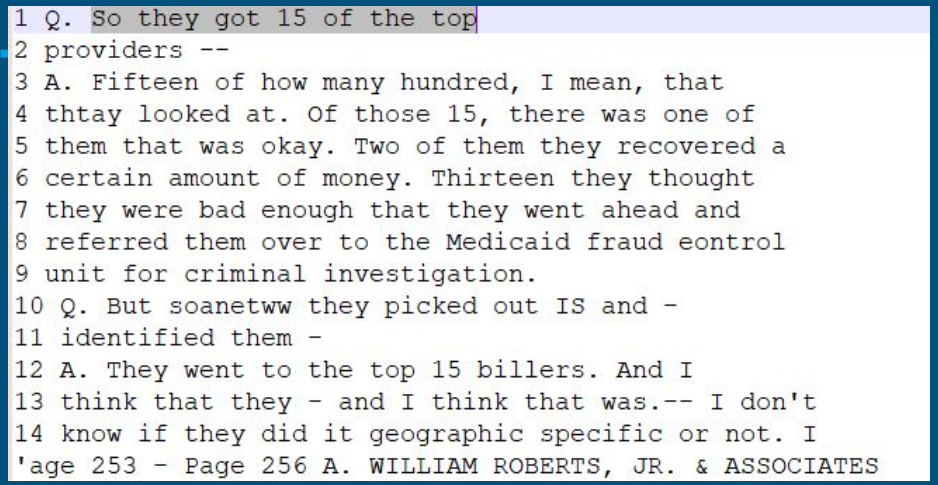
- Characters in documents which can represent noise for ML approaches (line numbers, page numbers, excessive spaces, etc.)
- Service symbols generated after OCR processing (\n, \r, etc.)
- Garbage characters appeared as a result of low quality of OCR (handwritten notes and OCR artifacts)

Document Text Cleaning - Algorithm

1. Read .ocr files from the collection one-by-one (using record keys)
2. Represent the file as a set of lines
3. Modify lines, deleting unnecessary characters
4. Reassemble and save file in a specified directory



Line Numbers Removal

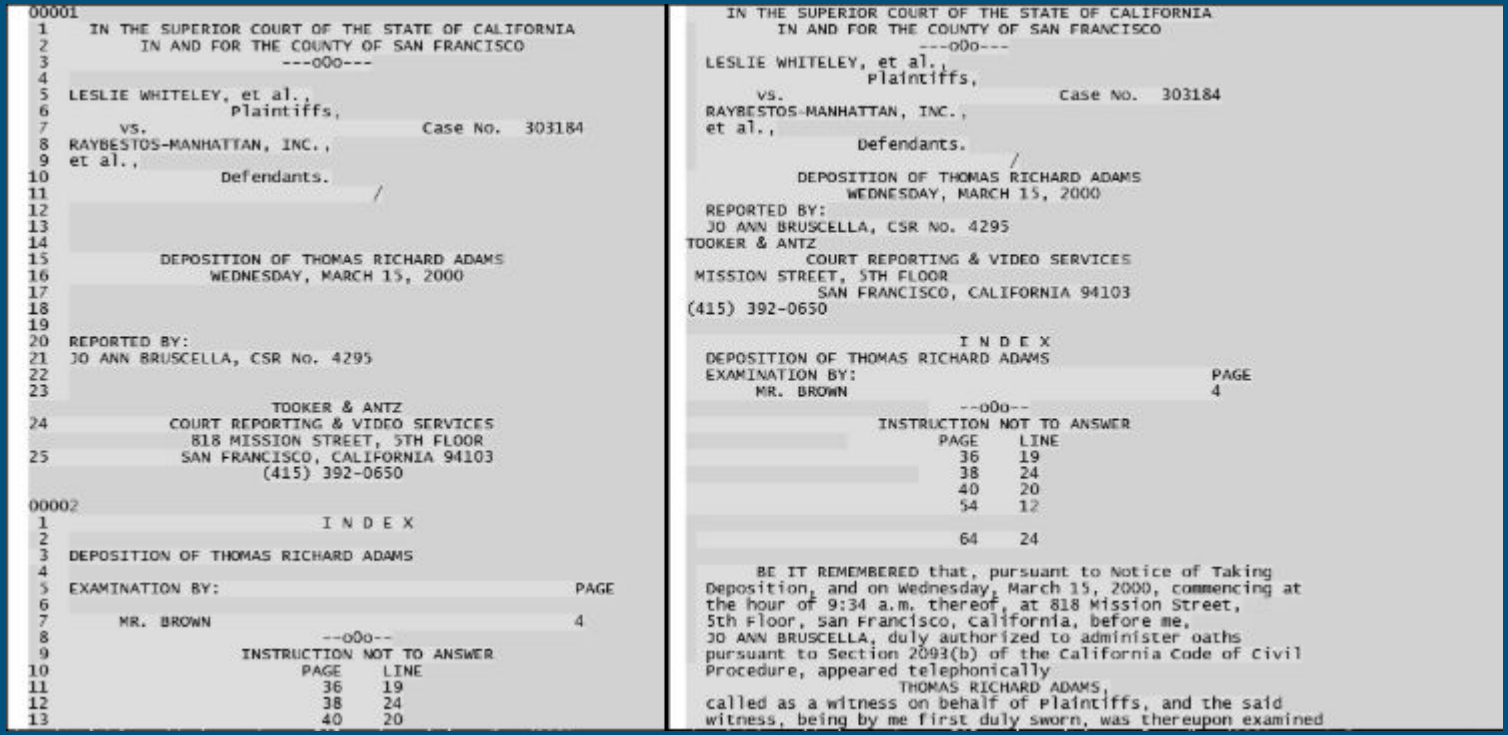


```
1 Q. So they got 15 of the top
2 providers --
3 A. Fifteen of how many hundred, I mean, that
4 thtay looked at. Of those 15, there was one of
5 them that was okay. Two of them they recovered a
6 certain amount of money. Thirteen they thought
7 they were bad enough that they went ahead and
8 referred them over to the Medicaid fraud eontrol
9 unit for criminal investigation.
10 Q. But soanetww they picked out IS and -
11 identified them -
12 A. They went to the top 15 billers. And I
13 think that they - and I think that was.-- I don't
14 know if they did it geographic specific or not. I
'age 253 - Page 256 A. WILLIAM ROBERTS, JR. & ASSOCIATES
```

Before
vs.
After

Q. So they got 15 of the top providers --
A. Fifteen of how many hundred, I mean, that thtay looked at. Of those 15, there was one of them that was okay. Two of them they recovered a certain amount of money. Thirteen they thought they were bad enough that they went ahead and referred them over to the Medicaid fraud eontrol unit for criminal investigation.
Q. But soanetww they picked out IS and - identified them -
A. They went to the top 15 billers. And I think that they - and I think that was.-- I don't know if they did it geographic specific or not. I WHIDDON, JOHN. Condenselt' FLORIDA vs. TOBACCO i would be what, you referring specifically to 1 problematic, and this is why it's going to be Unisys or historically here? 2 difficult to deal with in that regard.

Page and Line Numbers Removal (Before vs. After)



00001

IN THE SUPERIOR COURT OF THE STATE OF CALIFORNIA
IN AND FOR THE COUNTY OF SAN FRANCISCO
---o0o---

LESLIE WHITELEY, et al.,
Plaintiffs,
vs. Case No. 303184
RAYBESTOS-MANHATTAN, INC.,
et al.,
Defendants.
/

DEPOSITION OF THOMAS RICHARD ADAMS
WEDNESDAY, MARCH 15, 2000

REPORTED BY:
JO ANN BRUSCELLA, CSR No. 4295

TOOKER & ANTZ
COURT REPORTING & VIDEO SERVICES
818 MISSION STREET, 5TH FLOOR
SAN FRANCISCO, CALIFORNIA 94103
(415) 392-0650

00002

I N D E X

DEPOSITION OF THOMAS RICHARD ADAMS

EXAMINATION BY: PAGE

MR. BROWN 4

--o0o--

INSTRUCTION NOT TO ANSWER

PAGE	LINE
36	19
38	24
40	20

IN THE SUPERIOR COURT OF THE STATE OF CALIFORNIA
IN AND FOR THE COUNTY OF SAN FRANCISCO
---o0o---
LESLIE WHITELEY, et al.,
Plaintiffs,
vs. Case No. 303184
RAYBESTOS-MANHATTAN, INC.,
et al.,
Defendants.
/
DEPOSITION OF THOMAS RICHARD ADAMS
WEDNESDAY, MARCH 15, 2000
REPORTED BY:
JO ANN BRUSCELLA, CSR No. 4295
TOOKER & ANTZ
COURT REPORTING & VIDEO SERVICES
MISSION STREET, 5TH FLOOR
SAN FRANCISCO, CALIFORNIA 94103
(415) 392-0650
I N D E X
DEPOSITION OF THOMAS RICHARD ADAMS
EXAMINATION BY: PAGE
MR. BROWN 4
--o0o--
INSTRUCTION NOT TO ANSWER
PAGE LINE
36 19
38 24
40 20
54 12

64 24

BE IT REMEMBERED that, pursuant to Notice of Taking Deposition, and on Wednesday, March 15, 2000, commencing at the hour of 9:34 a.m. thereof, at 818 Mission Street, 5th Floor, San Francisco, California, before me, JO ANN BRUSCELLA, duly authorized to administer oaths pursuant to Section 2093(b) of the California Code of Civil Procedure, appeared telephonically

THOMAS RICHARD ADAMS,

called as a witness on behalf of Plaintiffs, and the said witness, being by me first duly sworn, was thereupon examined

Limitations: Line Numbers within Text

```
i   would be what, you referring specifically to    1   problematic, and this is why it's going to be
2   Unisys or historically here?    2   difficult to deal with in that regard.
3   A. Well, yeah. Well, historically, but I'm  3   Q. What's the triangle?
4   thinking Unisys.    4   A. I just said that was my - for me to kind
5   The thing that triggered my thought 5   of remember my thinking on number three there, the
6   here was I think it was the grand jury report that  6   trilogy with the kicker that I said.
7   I reviewed that they made a big issue about the 7   Q. Okay. To the right of the
8   hospital bids and the initiative that they  8   triangle, what are those words?
9   developed to try to go after the money in these 9   A. That was just my thinking about to some
10  areas. And what they discovered'in that process 10  extent for the trilogy that all claims require, in
11  was that a system edit that everybody thought was   11  order to deal with to some degree of health care
12  working that should have been tested in the 12  fraud and abuse, to get at it, you got to do a
13  implementation portion of the fiscal agent  13  substantive rt.wiew. To do a more intensive one,
14  contract obviously was not working, and therefore   14  you got to do a claim-by-claim review. And then
15  payments had continued to be made on these long 15  the third part is that even though you do a
16  after the real purchasing period was up and 16  claim-by-claim review, you may never know even
17  resulted in significant overpayments as a result    17  then.
1 s of that.    18  And then of course I think the
```

Garbage Characters Removal

```
5
~Aew.~eFl«,a.
~` 2'ao~a~~°°~ "'i° 5   Jersey and New York, do hereby certify that prior
6   and Insi~hts ..... not marked
~
'
~   6   to the COmmCfICCIIlent of the examination the witness
7   - m-- ~~
`
```

Before vs. After

Whiddon..... not marked 4 Reporter and Notary Public of the States of New 5 Jersey and New York, do hereby certify that prior and Insi~hts ..... not marked 6 to the COmmCfICCIIlent of the examination the witness - m-- ~~ 7 was sworn by me to testify the truth, the whole 8 truth and nothing but the truth. 9 I do fusther certify that the foregoing is 10 a true and accurate transcript of the testimony as I I 11 taken stenographically by and before me at the 12 tin>c, place and on the date hereinbefore set 13 forth. 14 I do further oertify that I am neithcr of is counsel nor attorney for any party in this action 16 and that I am not interested in the event nor Ln outCOlne of this litigation. . ~ 18 -J 19 m 20 w 21 ~ 22 ru. Iftakwoo2ass 23 N.Y. Registration No. 24-5006375 24 New Jersey Certificate No. 3Q01287 25 My N.J. commission expires December 11, 1997

Document Text Preprocessing - Overview

Some of ML algorithms require normalized text data instances.

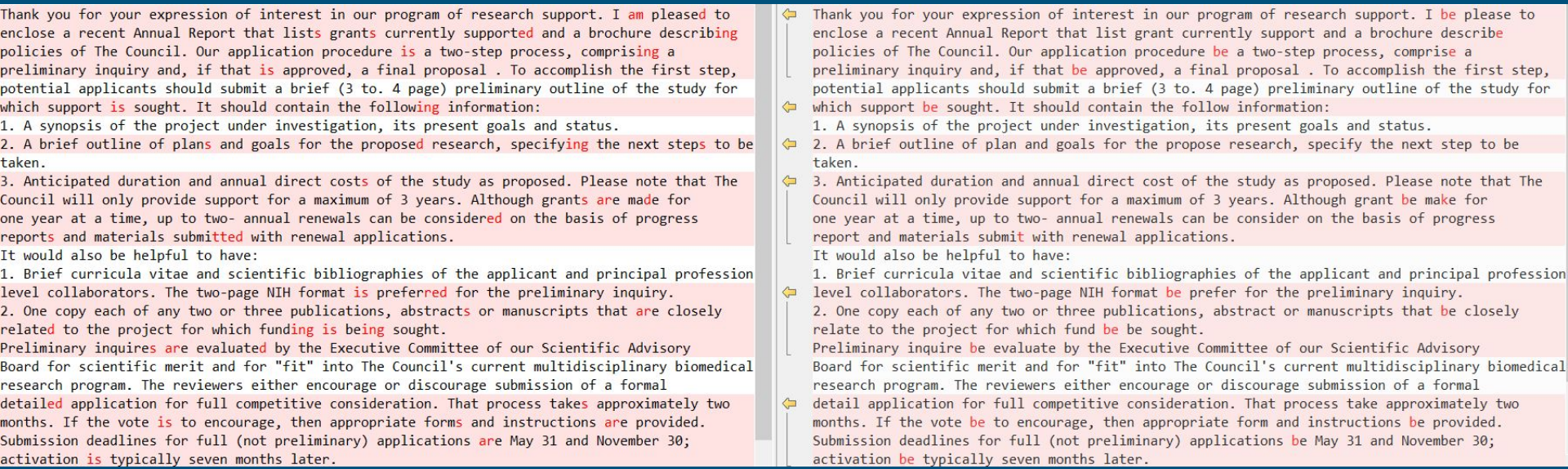
Text preprocessing includes:

- Tokenization (splitting texts into lexical entities - tokens)
- Lemmatization (bringing the tokens into their initial lexical form)
- Concatenation of lines in documents of ‘Q&A’ type (for instance, articles)

Results of Tokenization

```
['THE', 'COUNCIL', 'FOR', 'TOBACCO', 'RESEARCH-TT.S.A.', ',', 'INC.', '900', 'TIiIRD', 'AVENUE', 'NEW', 'YORK', '.', 'N.', 'Y', '.', '10022', 'December', '23', ',',
'1986', 'Dean', 'Befus', ',', 'Ph.D', '.', 'The', 'University', 'of', 'Calgary', 'Health', 'Sciences', 'Centre', '3330', 'Hospital', 'Drive', 'N.W', '.', 'Calgary',
',', 'Alberta', ',', 'Canada', 'T2N', '4N1', 'Dear', 'Dr.', 'Befus', ':', 'Thank', 'you', 'for', 'your', 'expression', 'of', 'interest', 'in', 'our', 'program', 'of',
'research', 'support', '.', 'I', 'am', 'pleased', 'to', 'enclose', 'a', 'recent', 'Annual', 'Report', 'that', 'lists', 'grants', 'currently', 'supported', 'and', 'a',
'brochure', 'describing', 'policies', 'of', 'The', 'Council', '.', 'Our', 'application', 'procedure', 'is', 'a', 'two-step', 'process', ',', 'comprising', 'a',
'preliminary', 'inquiry', 'and', ',', 'if', 'that', 'is', 'approved', ',', 'a', 'final', 'proposal', '.', 'To', 'accomplish', 'the', 'first', 'step', ',',
'potential', 'applicants', 'should', 'submit', 'a', 'brief', '(', '3', 'to', '.', '4', 'page', ')', 'preliminary', 'outline', 'of', 'the', 'study', 'for', 'which',
'support', 'is', 'sought', '.', 'It', 'should', 'contain', 'the', 'following', 'information', ':', '1', '.', 'A', 'synopsis']
```

Results of Lemmatization - Before vs. After



Thank you for your expression of interest in our program of research support. I am pleased to enclose a recent Annual Report that lists grants currently supported and a brochure describing policies of The Council. Our application procedure is a two-step process, comprising a preliminary inquiry and, if that is approved, a final proposal . To accomplish the first step, potential applicants should submit a brief (3 to. 4 page) preliminary outline of the study for which support is sought. It should contain the following information:

1. A synopsis of the project under investigation, its present goals and status.
2. A brief outline of plans and goals for the proposed research, specifying the next steps to be taken.
3. Anticipated duration and annual direct costs of the study as proposed. Please note that The Council will only provide support for a maximum of 3 years. Although grants are made for one year at a time, up to two- annual renewals can be considered on the basis of progress reports and materials submitted with renewal applications.

It would also be helpful to have:

1. Brief curricula vitae and scientific bibliographies of the applicant and principal profession level collaborators. The two-page NIH format is preferred for the preliminary inquiry.
2. One copy each of any two or three publications, abstracts or manuscripts that are closely related to the project for which funding is being sought.

Preliminary inquires are evaluated by the Executive Committee of our Scientific Advisory Board for scientific merit and for "fit" into The Council's current multidisciplinary biomedical research program. The reviewers either encourage or discourage submission of a formal detailed application for full competitive consideration. That process takes approximately two months. If the vote is to encourage, then appropriate forms and instructions are provided. Submission deadlines for full (not preliminary) applications are May 31 and November 30; activation is typically seven months later.

Thank you for your expression of interest in our program of research support. I be please to enclose a recent Annual Report that list grant currently support and a brochure describe policies of The Council. Our application procedure be a two-step process, comprise a preliminary inquiry and, if that be approved, a final proposal . To accomplish the first step, potential applicants should submit a brief (3 to. 4 page) preliminary outline of the study for which support be sought. It should contain the follow information:

1. A synopsis of the project under investigation, its present goals and status.
2. A brief outline of plan and goals for the propose research, specify the next step to be taken.
3. Anticipated duration and annual direct cost of the study as proposed. Please note that The Council will only provide support for a maximum of 3 years. Although grant be make for one year at a time, up to two- annual renewals can be consider on the basis of progress report and materials submit with renewal applications.

It would also be helpful to have:

1. Brief curricula vitae and scientific bibliographies of the applicant and principal profession level collaborators. The two-page NIH format be prefer for the preliminary inquiry.
2. One copy each of any two or three publications, abstract or manuscripts that be closely relate to the project for which fund be be sought.

Preliminary inquire be evaluate by the Executive Committee of our Scientific Advisory Board for scientific merit and for "fit" into The Council's current multidisciplinary biomedical research program. The reviewers either encourage or discourage submission of a formal detail application for full competitive consideration. That process take approximately two months. If the vote be to encourage, then appropriate form and instructions be provided. Submission deadlines for full (not preliminary) applications be May 31 and November 30; activation be typically seven months later.

Concatenation of Q's and A's - preliminary results

Q. Okay. I assume that the Congressional Record that incorporates the portions of Exhibit 2 are in the materials that were provided to me?
A. Yes, I have seen -- I beliove that they wen:, yes.

Before vs. After

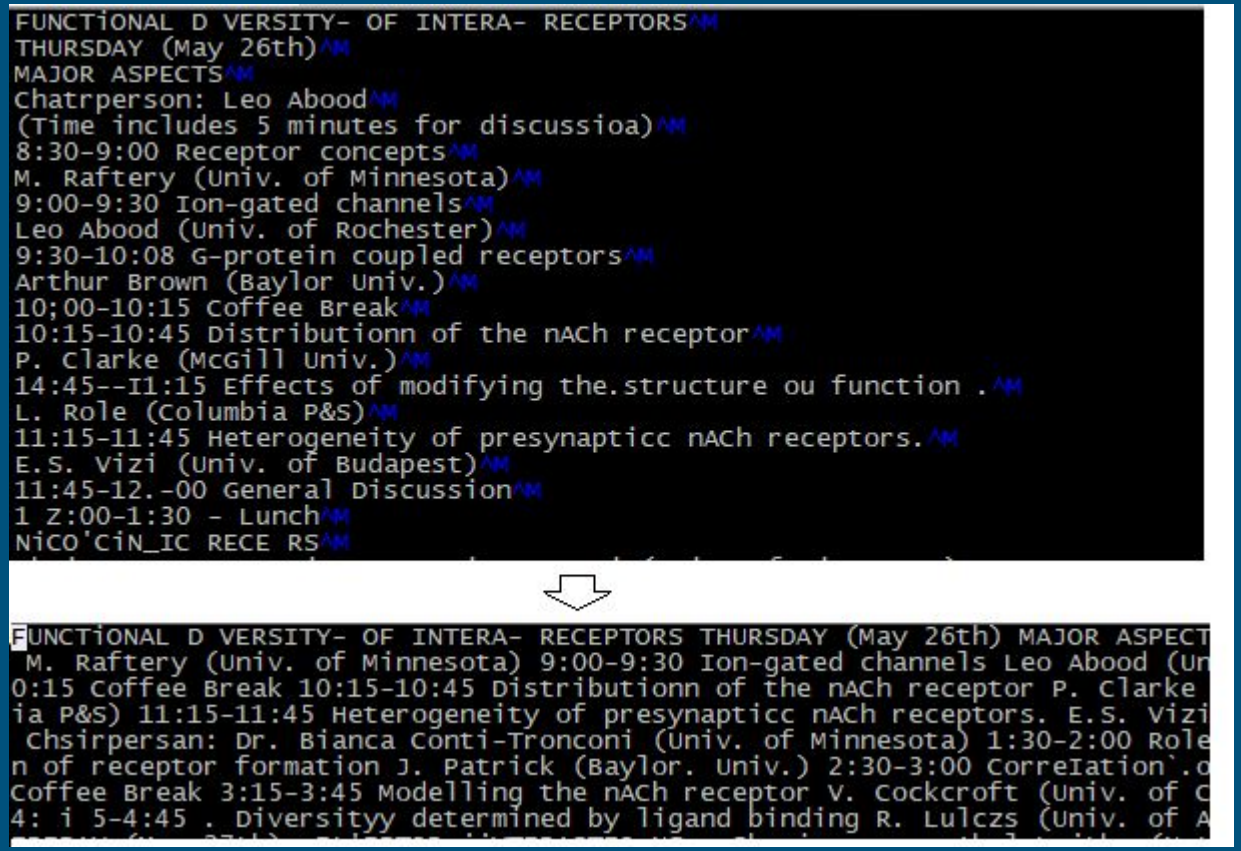
Q. Okay. I assume that the Congressional Record that incorporates the portions of Exhibit 2 are in the materials that were provided to me?
A. Yes, I have seen -- I beliove that they wen:, yes. MR. ULLMAN: For your information, I'm not - I'm not sure whetl>cr it's actually the Congressional Record as such as opposed to hearings before a subcommittee, but it's in an official U.S. governntent publication I believe that was sent to you. And I'm also going to give you at this time a copy of Mr. Whiddon's engagement letter, David. MR. FONVIELLE: 0ood. .

Format Adjustment - Overview

All the preprocessed text need to be transformed into the format that is digestible by the end system (used in search engine). The following steps are required:

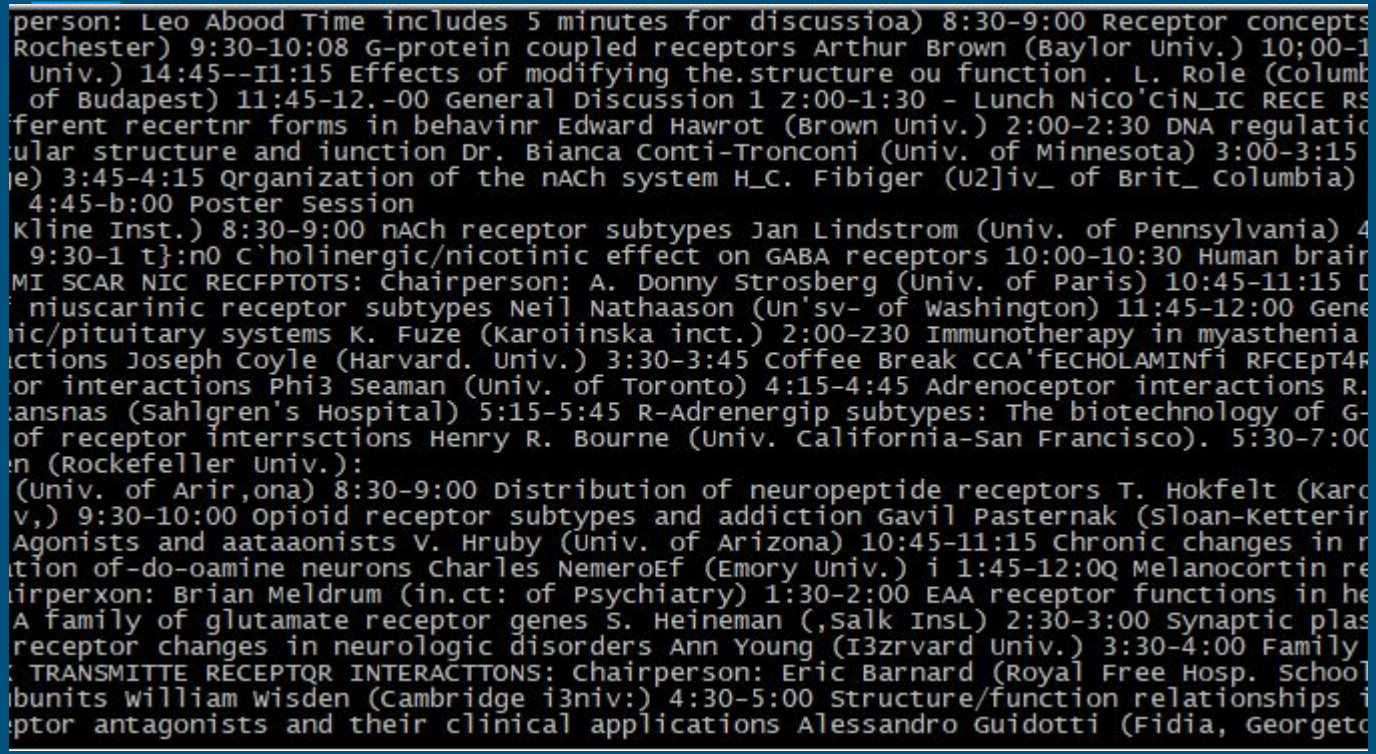
- Concatenate lines
- Concatenate pages
- Convert files to JSON format
- Add text field to metadata JSON file
- Extract data from OCR files

Concatenation



Before
vs.
After

Extract text from OCR files

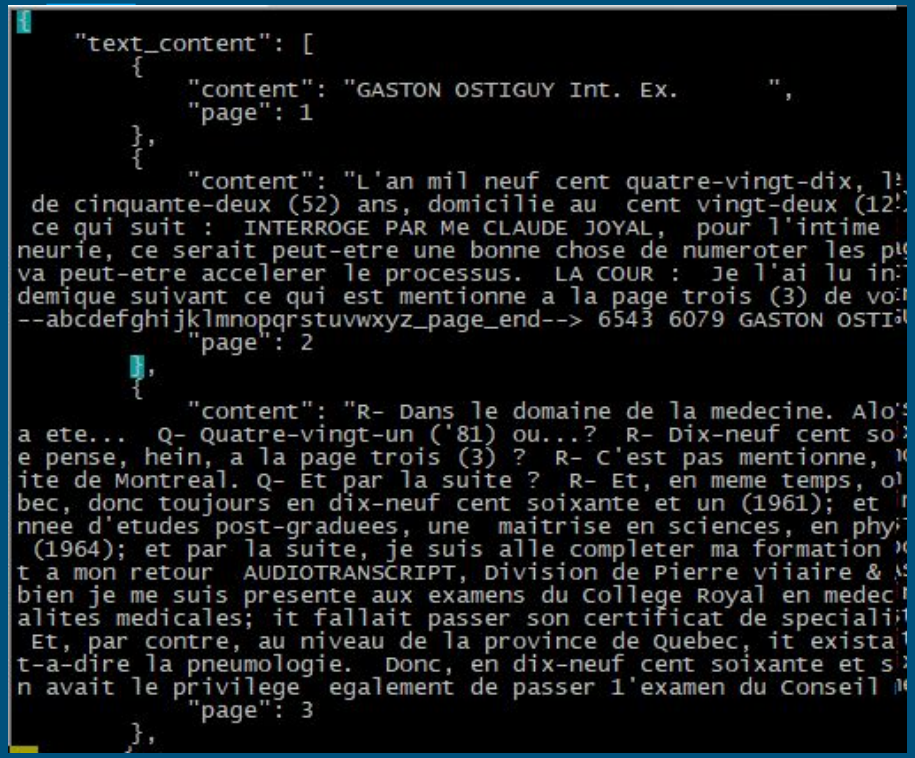


Page 1

Page 2

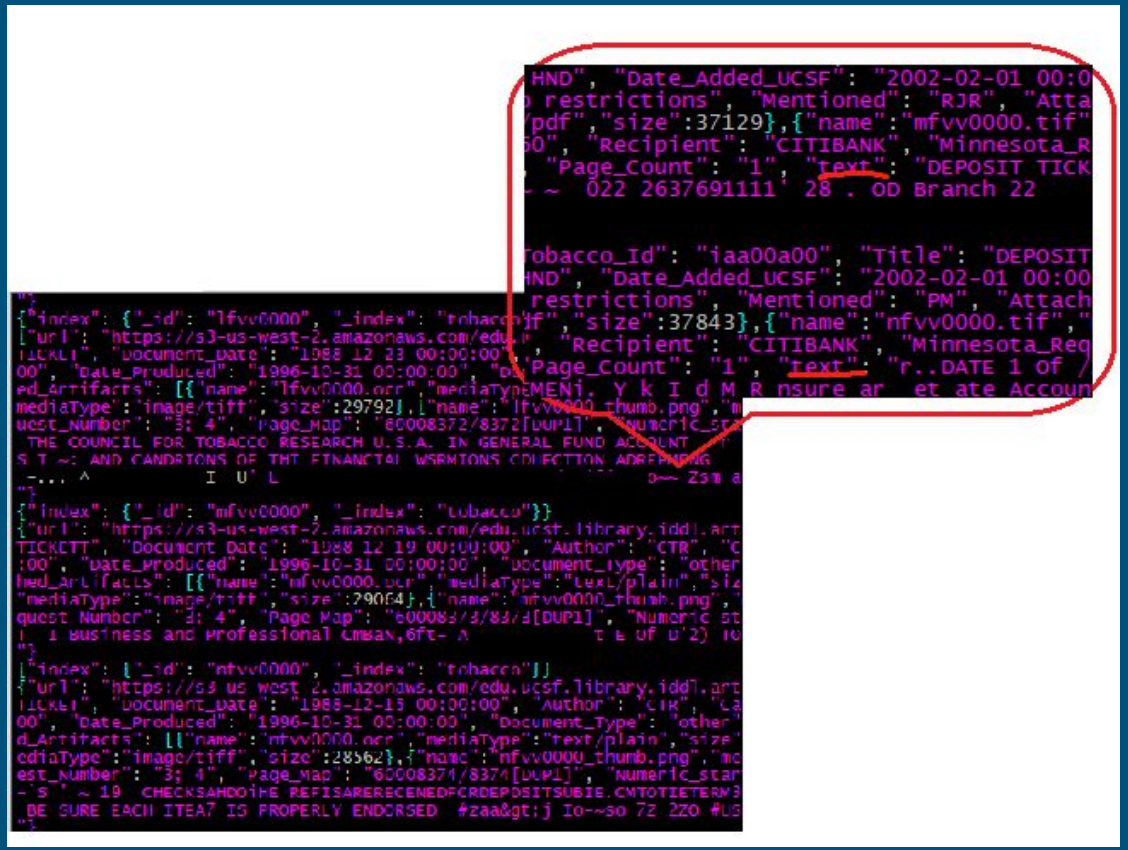
page3

Converting file to JSON format



“Text_content”:[
{“page”: 1,
“content”:”xxxx”},
{“page”: 2,
“content”:”xxxx”}]

Adding Text Field to Metadata JSON



“text”:”XXXX”

Text Processing - Deliverables

1. Text cleaning:
- Documents of ‘deposition’ type were cleaned (with line number removal)
- Millions of documents were cleaned (without line removal) and formatted (concatenated lines and pages)

1. Text preprocessing:
- A small subset of documents (depositions) was tokenized and lemmatized as requested by TML

1. Format adjustment:
- Depositions and articles files provided in JSON format
- Metadata with text fields for ELS Team

Unit Testing : JSON Validation

- Validation script written to check the contents of the JSON files provided to the ELS team.
- This ensures that the JSON provided will not fail at a later point during ingestion.

OCR Comparison

Motivation for OCR Comparison

- Documents OCR’ed with a proprietary tool by UCSF
  - Documents come in a variety of formats
  - Does not work well for documents with special format (e.g., multi-column, tables)
  - Creates garbage characters from handwritten text
- Alternative open-source OCR methods explored:
  - PyPDF2
  - PDFMiner
  - Abbyy Cloud OCR
- Goal
  - Compare different OCR methods across different document formats that represent the whole collection of documents

Performance on different types of documents

	UCSF	PyPDF2	PDFMiner	Abbyy Cloud OCR
Two Column PDFs	Poor to Average (comparable to pdfminer)	Poor	Average (comparable to UCSF)	Good
Newspaper Articles	Very poor	Extremely poor / No result	Very poor	Average
Tables	Average (comparable to pdfminer)	Poor	Average (comparable to UCSF)	Excellent
Plain texts, letters	Excellent	Good	Excellent	Excellent
Handwritten Texts	Very Poor	Extremely poor	Very Poor	Poor to Average

Original Documents

Document A: frlf0127

This memo is confidential to the business of the company: it should be carefully handled, is not transferable to another individual, and is not to be photocopied.

POL MEMO 0915

DATE: March 25, 1991

POL TEST NUMBER: 0915

TITLE: POL 0915 - Marlboro 85mm vs. Camel 85mm

TEST REQUESTER: B. Monahan

POL STUDY LEADER: A. L. Manwaring

DATA ANALYZED/RESULTS: A. L. Manwaring/D. Purvis

PREPARED BY: D. Purvis

PROCEDURE [BALLOT]: Sequential Monadic NV1BSN

SAMPLE SIZE: 800 **% RETURNED:** 83

SMOKER GROUPS:	**Usable Returns**
Marlboro 80/85mm	289
Camel 80/85mm	272

ANALYTICAL RESULTS: (See attachment)

CIGARETTES MADE: Marlboro 85mm 10/90
Camel 85mm 9/90

CIGARETTES SHIPPED: 1/15/91

DATA DATE: 2/21/91

RESULTS DISCUSSED WITH REQUESTER [DATE]: 2/22/91

RESULTS: There were significant differences between the cigarette ratings. The Marlboro 80/85mm smokers preferred the Marlboro 85mm and rated it higher on mild taste, good taste, satisfaction, cool smoking, good aftertaste, and liking. They rated the Camel 85mm higher on harshness and dry taste. The Camel 80/85mm smokers preferred the Camel 85mm.

COMMENTS: A multivariate analysis of variance (BMDP4V) was run on the data. This analysis collapses the ten scales into one overall rating to determine whether one cigarette (Marlboro 85mm or Camel 85mm) was rated significantly different from the other one. The results from this analysis showed that there was a significant cigarette-by-brand interaction (p=.0047). That is, the Marlboro smokers rated the Marlboro 85mm cigarettes higher than the Camel 85mm cigarettes, while the Camel smokers rated both cigarettes similarly.

Document B: gpny0069

WEEKLY MEMO
TOTAL

INDUSTRY	VOLUME IN BILLIONS 1996	1997	VOL CHG	% CHG
YEAR-TO-DATE	438.341	436.367	-1.973	-0.5
WEEKLY AVG.	9.326	9.364	0.038	0.4
45	9.143	8.910	-0.233	-2.6
46	9.160	9.432	0.272	3.0
47	9.790	10.145	0.355	3.6
48	8.020	8.547	0.527	6.6
LATEST 4 WEEKS	36.114	37.035	0.921	2.5

LORILLARD	1996	1997	VOL CHG	% CHG	SOM
YEAR-TO-DATE	37.421	38.600	1.179	3.2	8.85
WEEKLY AVG.	0.796	0.828	0.032	4.0	8.85
45	0.820	0.796	-0.024	-3.0	8.93
46	0.780	0.849	0.069	8.8	9.00
47	0.804	0.881	0.077	9.6	8.68
48	0.665	0.797	0.132	19.9	9.33
LATEST 4 WEEKS	3.068	3.323	0.254	8.3	8.97

PyPDF2

**This memo is confidential to the business of the company: it shouldbe carefully handled, is not transferable to another indivi
dual, and isnot to be photocopied.**POL MEMO 0915DATE: March 25, 1991POL TEST NUMBER: 0915TITLE: POL 0915 - Marlboro 85mm vs.'Ca
mel 85mmTEST REQUESTER: B. MonahanPOL STUDY LEADER: A. L. ManwaringDATA ANALYZED/RESULTS: A. L. Manwaring/D. PurvisPREPARED BY:
D. Purvis PFPROCEDURE [BALLOT]: Sequential Monadic NVIBSNSAMPLE SIZE: 800 -W RETURNED: 83SMOKER GROUPS: Usable ReturnsMarlboro 8
0/85mm 289Camel 80/85mm 272ANALYTICAL RESULTS: (See attachment)CIGARETTES MADE: Marlboro 85mm 10/90 CIGARETTES SHIPPED: 1/15/91C
amel 85mm 9/90 =DATA DATE: 2/21/91RESULTS DISCUSSED WITH REQUESTER [DATE]: 2/22/91RESULTS: There were significant differences be
tween the cigarette ratings.The Marlboro 80/85mm smokers preferred the Marlboro 85mm and rated it higheron mild taste, good tast
e, satisfaction, cool smoking, good aftertaste, andliking. They rated the Camel 85mm higher on harshness and dry taste. TheCamel
80/85mm smokers preferred the Camel 85mm.COMtMENTS: A multivariate analysis of variance (BMDP4V) was run on the data.This analy
sis collapses the ten scales into one overall rating to determinewhether one cigarette (Marlboro 85mm or Camel 85mm) was rated s
ignificantlydifferent from the other one. The results from this analysis showed thatthere was a significant cigarette-by-brand i
nteraction (p=.0047). That is,the Marlboro smokers rated the Marlboro 85mm cigarettes higher than the Camel85mm cigarettes, whil
e the Camel smokers rated both cigarettes similarly.Source: https://www.industrydocuments.ucsf.edu/docs/frlf0127The data from th

Document A: frlf0127

WEEKLY MEMOTOTALVOLUME IN BILLIONSINDUSTRY 1996 1997 VOL CHG •lo CHGYEAR-TO-DATE 438.341 436.367 -1.973 -0.5WE
EKLY AVG. 9.326 9.364 0.038 0.445 9.143 8.910 -0.233 -2.646 9.160 9.432 0.272 3.047 9.790 10.145 0.355 3.648 8
.020 8.547 0.527 6.6LATEST 4 WEEKS 36.114 37.035 0.921 2.5LORILLARD 1996 ,1997 VOL CHG% CHGYEAR-TO-DATE 37.421
38.600 1.179 3.2 8.85WEEKLY AVG. 0.796 0.828 0.032 4.0 8.8545 0.820 0.796 -0.024 -3.0 8.9346 0.780 0.849 0.06
9 8.8 9.0047 0.804 0.881 0.077 9.6 8.6848 0.665 0.797 0.132 19.9 9.33LATEST 4 WEEKS 3.068 3.323 0.254 8.3 8.97
Source: https://www.industrydocuments.ucsf.edu/docs/gpny0069WEEKLY MEMOFULL PRICEVOLUME (BILLIONS)FULL PRICE 1
996 1997 VOL CHG% CHGElYEAR-TO-DATE 315.798 318.265 2.467 0.8 72.94WEEKLY AVG. 6.719 6.830 0.111 1.6 72.9445 6
.522 6.371 -0.152 -2.3 71.5046 6.572 6.864 0.293 4.5 72.7847 7.071 7.520 0.449 6.4 74.1248 5.827 6.342 0.515 8
.8 74.20LATEST 4 WEEKS 25.992 27.098 1.105 4.3 73.17LORILLARD FP 1996 1997 VOL CHG % CHG ,S,_QM •1n~.1.I1YEAR-
TO-DATE 35.035 34.401 -0.635 -1.8 7.88 89.1WEEKLY AVG. 0.745 0.738 -0.007 -1.0 7.88 89.145 0.728 0.714 -0.014
-1.9 8.02 89.846 0.714 0.760 0.046 6.5 8.06 89.547 0.743 0.787 0.044 5.9 7.75 89.348 0.615 0.719 0.103 16.8 8.
41 90.1LATEST 4 WEEKS 2.800 2.980 0.180 6.4 8.05 89.7Source: https://www.industrydocuments.ucsf.edu/docs/gpny0
069WEEKLY MEMODISCOUNTVOLUME (BILLIONS)DISCOUNT 1996 1997 VOL CHG % CHG SOMYEAR-TO-DATE 122.542 118.103 -4.440

Document B: gpny0069

Document A: frlf0127

UCSF

<ocr>**This memo is confidential to the business of the company: it should be carefully handled, is not transferable to another individual, and is not to be photocopied.**
POL MEMO 0915
DATE: March 25, 1991
POL TEST NUMBER: 0915
TITLE: POL 0915 - Marlboro 85mm vs.'Camel 85mm
TEST REQUESTER: B. Monahan
POL STUDY LEADER: A. L. Manwaring
DATA ANALYZED/RESULTS: A. L. Manwaring/D. Purvis
PREPARED BY: D. Purvis PF
PROCEDURE [BALLOT]: Sequential Monadic NVIBSN
SAMPLE SIZE: 800 -W RETURNED: 83
SMOKER GROUPS: Usable Returns
Marlboro 80/85mm 289
Camel 80/85mm 272
ANALYTICAL RESULTS: (See attachment)
CIGARETTES MADE: Marlboro 85mm 10/90 CIGARETTES SHIPPED: 1/15/91
Camel 85mm 9/90 =
DATA DATE: 2/21/91
RESULTS DISCUSSED WITH REQUESTER [DATE]: 2/22/91
RESULTS: There were significant differences between the cigarette ratings. The Marlboro 80/85mm smokers preferred the Marlboro 85mm and rated it higher on mild taste, good taste, satisfaction, cool smoking, good aftertaste, and liking. They rated the Camel 85mm higher on harshness and dry taste. The Camel 80/85mm smokers preferred the Camel 85mm.
COMtMENTS: A multivariate analysis of variance (BMDP4V) was run on the data. This analysis collapses the ten scales into one overall rating to determine whether one cigarette (Marlboro 85mm or Camel 85mm) was rated significantly different from the other one. The results from this analysis showed that there was a significant cigarette-by-brand interaction (p=.0047). That is, the Marlboro smokers rated the Marlboro 85mm cigarettes higher than the Camel 85mm cigarettes, while the Camel smokers rated both cigarettes similarly.
pgNbr=1
The data from this test was also compared with data from previous POLS* testing Marlboro 85mm versus Camel 85mm using an analysis of variance with test, presentation, and brand as grouping variables. The results showed significant cigarette-by-brand interactions on nine out of the ten scales. Generally, the cigarette-by-brand interactions showed that the Marlboro smokers rated the Marlboro cigarettes more favorably than the Camel

PDFMiner

**This memo is confidential to the business of the company : it should be carefully handled, is not transferable to another individual, and is

not to be photocopied .**

POL MEMO 0915

DATE : March 25, 1991

POL TEST NUMBER : 0915

TITLE : POL 0915 - Marlboro 85mm vs .'Camel 85mm

TEST REQUESTER : B . Monahan

POL STUDY LEADER : A . L . Manwaring

DATA ANALYZED/RESULTS : A . L . Manwaring/D . Purvis

PREPARED BY : D . Purvis PF

PROCEDURE [BALLOT] : Sequential Monadic NVIBSN

SAMPLE SIZE : 800 -W RETURNED : 83

SMOKER GROUPS : Usable Returns

Marlboro 80/85mm 289
Camel 80/85mm 272

ANALYTICAL RESULTS : (See attachment)

CIGARETTES MADE : Marlboro 85mm 10/90 CIGARETTES SHIPPED : 1/15/91

Camel 85mm 9/90 =

DATA DATE : 2/21/91

RESULTS DISCUSSED WITH REQUESTER [DATE] : 2/22/91

RESULTS : There were significant differences between the cigarette ratings . The Marlboro 80/85mm smokers preferred the Marlboro 85mm and rated it higher on mild taste, good taste, satisfaction, cool smoking, good aftertaste, and

Document B: gpny0069

UCSF

```
WEEKLY MEMO
TOTAL
VOLUME IN BILLIONS
INDUSTRY 1996 1997 VOL CHG <B0>lo CHG
YEAR-TO-DATE 438.341 436.367 -1.973 -0.5
WEEKLY AVG. 9.326 9.364 0.038 0.4
45 9.143 8.910 -0.233 -2.6
46 9.160 9.432 0.272 3.0
47 9.790 10.145 0.355 3.6
48 8.020 8.547 0.527 6.6
LATEST 4 WEEKS 36.114 37.035 0.921 2.5
LORILLARD 1996 ,1997 VOL CHG
% CHG
YEAR-TO-DATE 37.421 38.600 1.179 3.2 8.85
WEEKLY AVG. 0.796 0.828 0.032 4.0 8.85
45 0.820 0.796 -0.024 -3.0 8.93
46 0.780 0.849 0.069 8.8 9.00
47 0.804 0.881 0.077 9.6 8.68
48 0.665 0.797 0.132 19.9 9.33
LATEST 4 WEEKS 3.068 3.323 0.254 8.3 8.97
pgNbr=1
WEEKLY MEMO
FULL PRICE
VOLUME (BILLIONS)
FULL PRICE 1996 1997 VOL CHG
% CHG
El
YEAR-TO-DATE 315.798 318.265 2.467 0.8 72.94
WEEKLY AVG. 6.719 6.830 0.111 1.6 72.94
45 6.522 6.371 -0.152 -2.3 71.50
46 6.572 6.864 0.293 4.5 72.78
47 7.071 7.520 0.449 6.4 74.12
48 5.827 6.342 0.515 8.8 74.20
LATEST 4 WEEKS 25.992 27.098 1.105 4.3 73.17
```

PDFMiner

```
WEEKLY MEMO

TOTAL

VOLUME IN BILLIONS

INDUSTRY 1996 1997 VOL CHG °lo CHG

YEAR-TO-DATE 438 .341 436 .367 -1 .973 -0 .5
WEEKLY AVG. 9.326 9 .364 0.038 0 .4

45 9 .143 8.910 -0.233 -2 .6
46 9 .160 9.432 0.272 3 .0
47 9 .790 10.145 0.355 3 .6
48 8 .020 8 .547 0.527 6 .6

LATEST 4 WEEKS 36.114 37 .035 0.921 2 .5

LORILLARD 1996 ,1997 VOL CHG

% CHG

YEAR-TO-DATE 37 .421 38 .600 1 .179 3 .2 8.85
WEEKLY AVG. 0.796 0 .828 0.032 4.0 8.85

45 0.820 0.796 -0 .024 -3.0 8.93
46 0 .780 0.849 0.069 8.8 9.00
47 0 .804 0 .881 0.077 9.6 8.68
48 0 .665 0 .797 0 .132 19.9 9.33

LATEST 4 WEEKS 3 .068 3 .323 0.254 8 .3 8.97

Source: https://www.industrydocuments.ucsf.edu/docs/gpny0069
```

Abbyy Cloud OCR

Document A: frlf0127

<U+FEFF>**This memo is confidential to the business of the company: it should be carefully handled, is not transferable to another individual, and is not to be photocopied.**

POL MEMO 0915

DATE: March 25, 1991
POL TEST NUMBER: 0915

TITLE: POL 0915 - Marlboro 85mm vs. Camel 85mm
TEST REQUESTER: B. Monahan
POL STUDY LEADER: A. L. Manwaring
DATA ANALYZED/RESULTS: A. L. Manwaring/D. Purvis
PREPARED BY: D. Purvis
PROCEDURE [BALLOT]: Sequential Monadic NV1BSN
SAMPLE SIZE: 800 ' % RETURNED: 83
SMOKER GROUPS: Usable Returns
Marlboro 80/85mm 289
Camel 80/85mm 272
ANALYTICAL RESULTS: (See attachment) ...
CIGARETTES MADE: Marlboro 85mm 10/90 CIGARETTES SHIPPED: 1/15/91
Camel 85mm 9/90 1
DATA DATE: 2/21/91 „
RESULTS DISCUSSED WITH REQUESTER [DATE]: 2/22/91 .

RESULTS: There were significant differences between the cigarette ratings. The Marlboro 80/85mm smokers preferred the Marlboro 85mm and rated it higher on mild taste, good taste, satisfaction, cool smoking, good aftertaste, and liking. They rated the Camel 85mm higher on harshness and dry taste. The Camel 80/85mm smokers preferred the Camel 85mm.
COMMENTS: A multivariate analysis of variance (BMDP4V) was run on the data. This analysis collapses the ten scales into one overall rating to determine whether one cigarette (Marlboro 85mm or Camel 85mm) was rated significantly different from the other one. The results from this analysis showed that there was a significant cigarette-by-brand interaction (p=.0047). That is, the Marlboro smokers rated the Marlboro 85mm cigarettes higher than the Camel 85mm cigarettes, while the Camel smokers rated both cigarettes similarly.

Document B: gpny0069

<U+FEFF>

WEEKLY MEMO

TOTAL

VOLUME IN BILLIONS

INDUSTRY	1W	199Z	VOLCHG	
YEAR-TO-DATE	438.341	436.367	-1.973	-0.5
WEEKLY AVG.	9.326	9.364	0.038	0.4
45	9.143	8.910	-0.233	-2.6
46	9.160	9.432	0.272	3.0
47	9.790	10.145	0.355	3.6
48	8.020	8.547	0.527	6.6
LATEST 4 WEEKS	36.114	37.035	0.921	2.5

LORILLARP	1996	1997	VOLCHG	% CHG	SOM
YEAR-TO-DATE	37.421	38.600	1.179	3.2	8.85
WEEKLY AVG.	0.796	0.828	0.032	4.0	8.85
45	0.820	0.796	-0.024	-3.0	8.93
46	0.780	0.849	0.069	8.8	9.00
47	0.804	0.881	0.077	9.6	8.68
48	0.665	0.797	0.132	19.9	9.33
LATEST 4 WEEKS	3.068	3.323	0.254	8.3	8.97

Performance Based on Time & Cost

	UCSF	PyPDF2	PDFMiner	Abbyy Cloud OCR
Time	Unknown	Good	Good	Average (depending upon server speed)
Cost	Unknown	None	None	Paid

VM independent Metadata & Data retrieval for INT

Approach

Problem: Addition of new tobacco files to the system in the future.

Constraint:

- Inaccessibility of the database hosted on the tobacco.cs.vt.edu from the Computer Science Cluster used by the Integration team.
- Inability to set up Kafka in time.

Interim Work Around:

- Creation of a demo folder outside the VM containing instances of metadata files, PDFs & text editable content of tobacco documents.
- Python script to pick metadata and data from these files and parse them into JSON format matching the one for tobacco files currently present on the VM.

Implementation

- Choose a set of dummy Tobacco Files:
  - Retrieve their metadata from the current DB and pull into a CSV file on local machine
- Convert metadata to JSON:
  - Build Python script to-
    - Merge the CSV files of the tables idl_doc_tobacco and idl_doc_field based on common *ID* values; and export results into new CSV file 'output.csv'.
    - Store itag descriptions in an array 'itag_array'
    - Pick data from the columns of 'output.csv' that contain the *Itag* and *Value* associated with a document ID; retrieve corresponding itag description based on *itag_array*.
    - Parse the data into desired JSON format and store into a new file 'output.json'.
- Convert Data to JSON:
  - Build Python script to-
    - Download PDFs corresponding to each record key and convert them into text editable files using Optical Character Recognition (OCR).
    - Parse the contents of each text document into JSON format and store in separate files.

idl_doc_tobacco.csv

A	B	C	D	E	F	G	H
id	record_key	bates	collection_id	dm	document_category	pages	industry_id
0	ffvv0000	60008366-60008366[DUP1]	11	20020201	PUBLIC	1	2
1	gfvv0000	60008367-60008367[DUP1]	11	20020201	PUBLIC	1	2
2	hfvv0000	60008368-60008368[DUP1]	11	20020201	PUBLIC	1	2
3	xfvv0000	60008369-60008369[DUP1]	11	20020201	PUBLIC	1	2
4	jfvv0000	60008370-60008370[DUP1]	11	20020201	PUBLIC	1	2
5	kfvv0000	60008371-60008371[DUP1]	11	20020201	PUBLIC	1	2
6	lfvv0000	60008372-60008372[DUP1]	11	20020201	PUBLIC	1	2
7	mfvv0000	60008373-60008373[DUP1]	11	20020201	PUBLIC	1	2
8	nfvv0000	60008374-60008374[DUP1]	11	20020201	PUBLIC	1	2

idl_doc_field.csv

id	itag	value
10	1	kaa00a00
10	3	DEPOSIT TICKET
10	4	8-Dec-88
10	5	CTR
10	24	MNAG
10	25	RECORDS TRANSACTION; MAR
10	27	1-Feb-02
10	28	31-Oct-96
10	29	other
10	32	public
10	33	no restrictions
10	38	LOR
10	46	":"text/plain","size":530},{"name":"pfvv0000.pdf","mediaType":"applicati
10	100	1
11	1	laa00a00
11	4	21-Nov-88

Excerpt from Array containing metadata field descriptions

```
itag_desc = {
"0":    "IGNORE",
"1":    "Legacy_(LTDL2)_Tobacco_Id",
"2":    "Collection",
"3":    "Title",
"4":    "Document_Date",
"5":    "Author",
"6":    "Recommend",
"7":    "Organization_Author",
"8":    "Person_Author",
"10":   "Referenced_Document",
"11":   "Attending",
"12":   "Organization_Attending",
"13":   "Person_Attending",
"14":   "Brands",
"15":   "Alternate_Bates_Range",
"16":   "Bates_Mater",
"17":   "Other_Bates_Range",
"18":   "Bates_Number",
"19":   "Copied",
"20":   "Organization_Copied",
"21":   "Date_Added_Industry_Site",
"24":   "Case",
"25":   "Description",
"27":   "Date_Added_UCSF",
"28":   "Date_Produced",
"29":   "Document_Type",
"32":   "availablility",
"33":   "availablilitystatus",
"35":   "File_Number",
"36":   "Grant_Number",
"38":   "Mentioned",
"39":   "Organization_Mentioned"
```

output.csv

id	record_key	bates	collection_id	dm	document_category	pages	industry_id	itag	value
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	1	kaa00a00
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	3	DEPOSIT TICKET
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	4	32485
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	5	CTR
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	24	MNAG
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	25	RECORDS TRANSACTION; MAR
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	27	37288
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	28	35369
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	29	other
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	32	public
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	33	no restrictions
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	96	60008376
10	pfvv0000	60008376-60008376[DUP1]	11	20020201	PUBLIC	1	2	100	1
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	1	laa00a00
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	4	32468
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	5	STORR HG, CTR
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	19	BROWN JC;MULLEN CH
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	24	MNAG
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	25	REQUESTS MEMBERSHIP DUES
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	27	37288
11	qfvv0000	60008377-60008377[DUP1]	11	20020201	PUBLIC	1	2	28	35369

CSV file obtained after merging idl_doc_tobacco.csv and idl_doc_field.csv

JSON file containing metadata


```
{"index": {"_index": "tobacco", "_id": "pfvv0000"}}
{"url": "https://s3-us-west-2.amazonaws.com/edu.ucsf.library.iddl.artifacts/p/f/v/v/pfvv0000/pfvv0000.pdf", "Legacy_
(LTDL2)_Tobacco_Id": "kaa00a00", "Title": "DEPOSIT TICKET", "Document_Date": "1988-12-08 00:00:00", "Author": "CTR", "Case":
"MNAG", "Description": "RECORDS TRANSACTION; MAR", "Date_Added_UCSF": "2002-02-01 00:00:00", "Date_Produced": "1996-10-31
00:00:00", "Document_Type": "other", "availablility": "public", "availablilitystatus": "no restrictions", "Mentioned": "LOR",
"Attached_Artifacts": [{"name":"pfvv0000.ocr","mediaType":"text/plain","size":530},
{"name":"pfvv0000.pdf","mediaType":"application/pdf","size":31211},
{"name":"pfvv0000.tif","mediaType":"image/tiff","size":22006},
{"name":"pfvv0000_thumb.png","mediaType":"image/png","size":41539}], "Box_Number": "260", "Recipient": "CITIBANK",
"Minnesota_Request_Number": "3; 4", "Page_Map": "60008376/8376[DUP1]", "Numeric_start_bates": "60008376", "Numeric_end_bates":
"60008376", "Page_Count": "1"}
{"index": {"_index": "tobacco", "_id": "qfvv0000"}}
{"url": "https://s3-us-west-2.amazonaws.com/edu.ucsf.library.iddl.artifacts/q/f/v/v/qfvv0000/qfvv0000.pdf", "Legacy_
(LTDL2)_Tobacco_Id": "laa00a00", "Document_Date": "1988-11-21 00:00:00", "Author": "STORR HG, CTR", "Copied": "BROWN JC;MULLEN
CH", "Case": "MNAG", "Description": "REQUESTS MEMBERSHIP DUES", "Date_Added_UCSF": "2002-02-01 00:00:00", "Date_Produced":
"1996-10-31 00:00:00", "Document_Type": "letter", "availablility": "public", "availablilitystatus": "no restrictions",
"Attached_Artifacts": [{"name":"qfvv0000.ocr","mediaType":"text/plain","size":823},
{"name":"qfvv0000.pdf","mediaType":"application/pdf","size":29960},
{"name":"qfvv0000.tif","mediaType":"image/tiff","size":20790},
{"name":"qfvv0000_thumb.png","mediaType":"image/png","size":39190}], "Box_Number": "260", "Recipient": "RANDOUR PA, AMER
BRANDS", "Minnesota_Request_Number": "3; 4", "Page_Map": "60008377/8377[DUP1]", "Numeric_start_bates": "60008377",
"Numeric_end_bates": "60008377", "Page_Count": "1"}
{"index": {"_index": "tobacco", "_id": "rfvv0000"}}
```

Future Work

- Improve the text cleaning process (account for line numbers within the text and more precise identifying of garbage characters)
- Add more test coverage for the scripts
- Finalize the OCR method and apply it to the documents

Thank you

References

- Truth tobacco industry documents - https://www.industrydocuments.ucsf.edu/tobacco/
- Tobacco Tactics - https://tobaccotactics.org/index.php/Advertising_Strategy#cite_note-1
- Tobacco Industry Marketing -Fast Facts and Fact Sheets: Tobacco Industry Marketing
- Abbyy OCR SDK- https://www.ocrsdk.com/

Questions?