Team Text Analytics and Machine Learning (TML)

CS 5604 Information Storage and Retrieval (Fall 2019)

Instructor: Dr. Edward Fox (fox@vt.edu)
TA: Ziqian Song (ziqian@vt.edu)

Virginia Tech
Blacksburg, VA - 24061

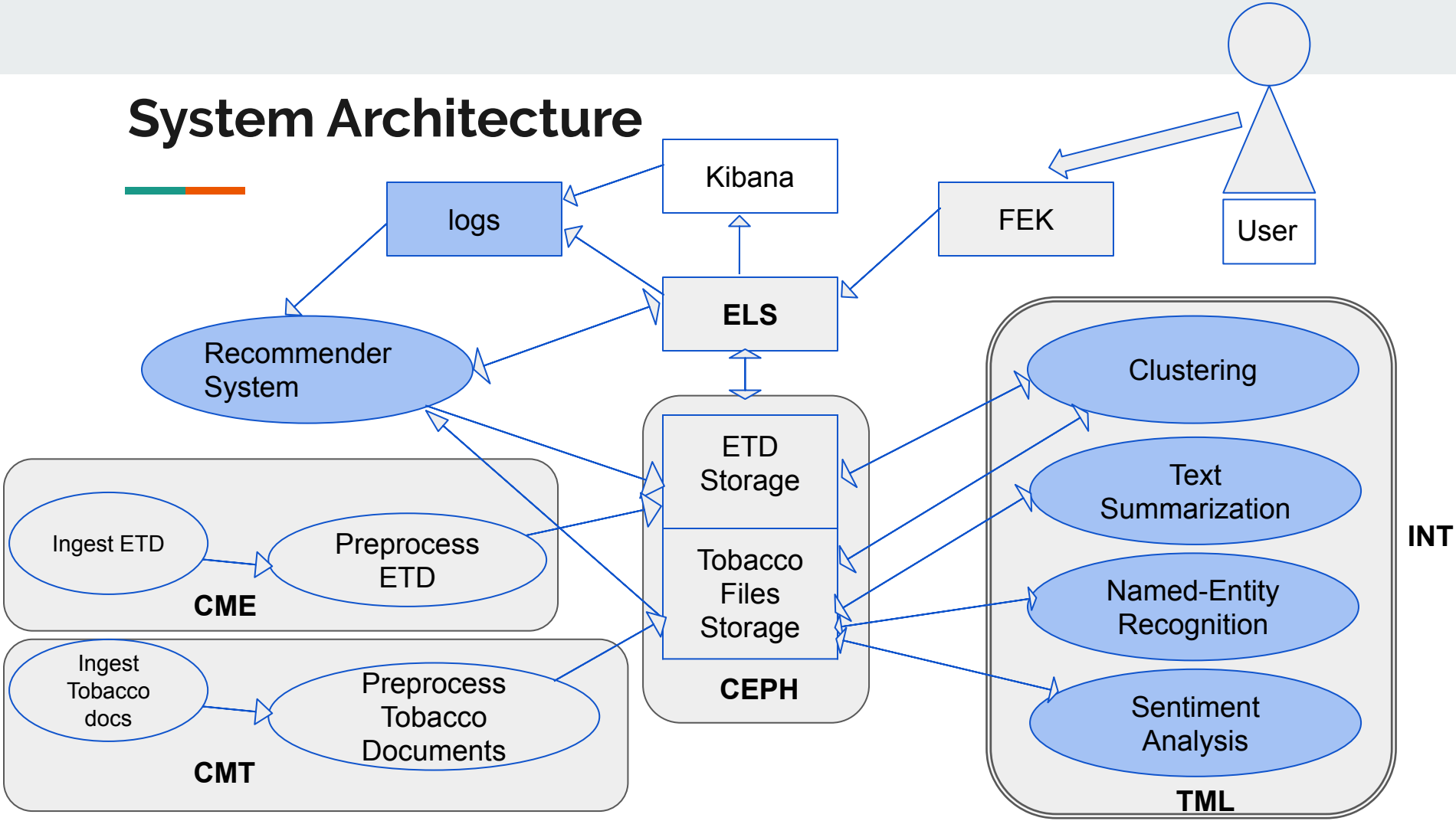
December 05, 2019

Adheesh Sunil Juvekar (juvekaradheesh@vt.edu)
Jiaying Gong (gjiaying@vt.edu)
Prathamesh Mandke (pkmandke@vt.edu)
Rifat Sabbir Mansur (rifatsm@vt.edu)
Sandhya M Bharadwaj (sandhyamb@vt.edu)
Sharvari Chougule (sharvarisc@vt.edu)

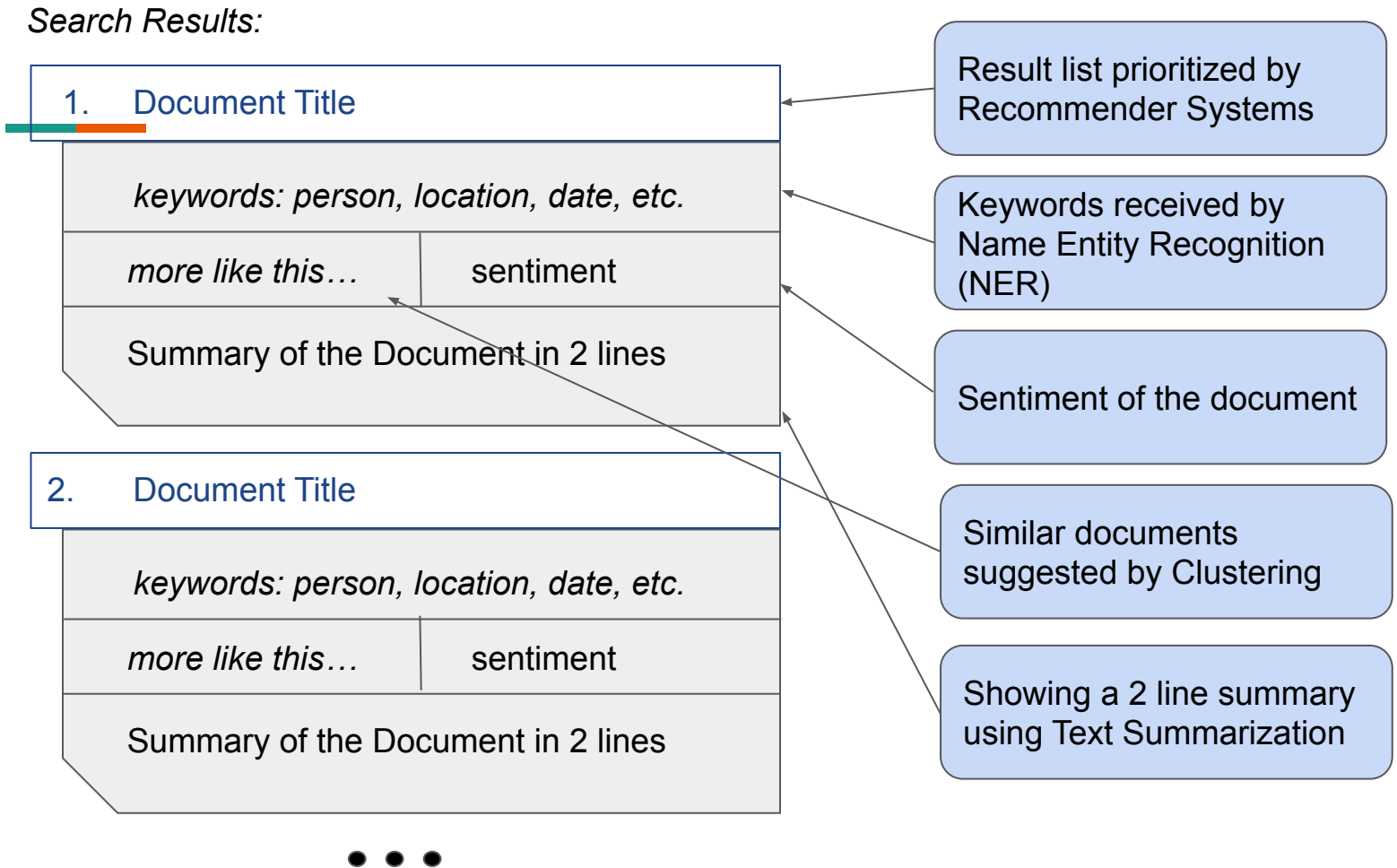
Outline

1. System Overview
2. Clustering
3. Text Summarization
4. Named Entity Recognition
5. Sentiment Analysis
6. Recommender Systems

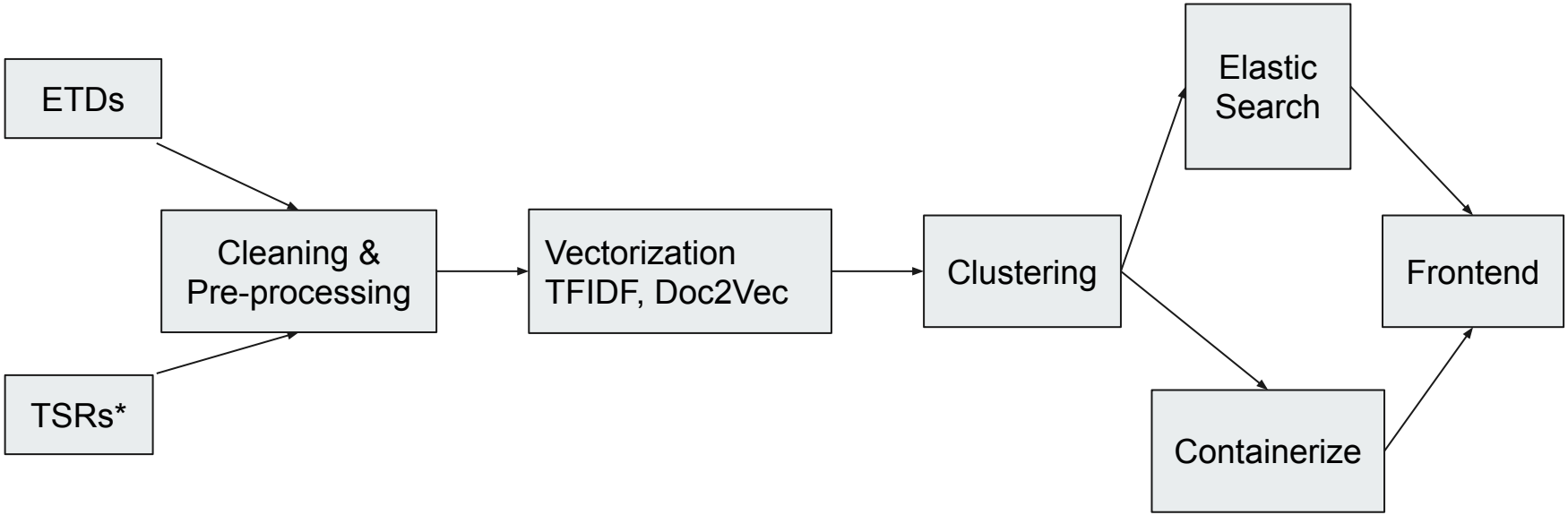
System Architecture



System Diagram



Clustering Workflow - A bird's eye-view!



*TSRs: Tobacco Settlement Records

Doc2Vec: Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. 2014.

Pre-Processing

- Step 1: Cleansing
  - Remove invalid UTF-8 characters
  - Remove punctuation and convert all characters/letters to lowercase
- Step 2: Tokenization
  - Punkt Sentence Tokenizer
  - Treebank Word Tokenizer
- Step 3: Stemming
  - Porter Stemmer

We also removed common English stopwords from the tokens

NLTK: https://www.nltk.org/api/nltk.chunk.html

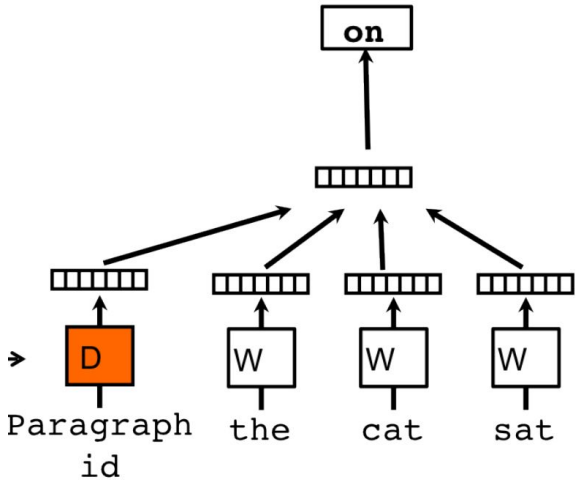
Clustering TSRs with TF-IDF vectors

- **Issues**
  - Imbalanced cluster allocation.
  - Results not very interpretable.
- **Possible causes**
  - Uncleaned/raw documents containing invalid characters -> leading to noisy TF-IDF vectors.
  - Highly sparse TF-IDF representations with only ~1% values in a vector being non-zero.
- **Result**
  - Unable to get good results with either K-Means or Agglomerative Clustering.

Cluster #	# of Documents
1	94
2	107
3	4806
4	283
5	283
6	320
7	340
8	123
9	529
10	259

Doc2Vec [1]

- From relative frequency counts to distributed representations that capture semantics.
- **Specifics**
  - 128-d document vectors for a total of 30961 ETDs.
  - Abstracts used to generate the document vectors.
  - Model trained for 15 epochs in a distributed memory setting using 5 parallel threads for data fetching on the ECE Guacamole servers.
- **Why 128-d vectors?**
  - Neither too big, nor to small!
  - Conducive to GPU implementations of downstream tasks that can use these document vectors.
  - Enough to capture information/semantics from the abstracts, entire documents will require higher dimensional vectors.



[1] Doc2Vec: Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. 2014. (Image source)

Outline of Clustering Experiments

1. K-Means Clustering
2. Hierarchical - Agglomerative Clustering
3. DBSCAN
4. BIRCH

Summary of Data Corpora

Corpus	Number of documents	Brief description
Uncleaned TSRs set	7995	Invalid bytes have not been removed.
Cleaned TSRs sample set	4553	All files have valid UTF-8 bytes.
Cleaned articles from TSRs	916977	Cleaned articles from the Tobacco corpus.
ETDs all	30961 (13071D + 17890T)	Text and metadata for all ETDs

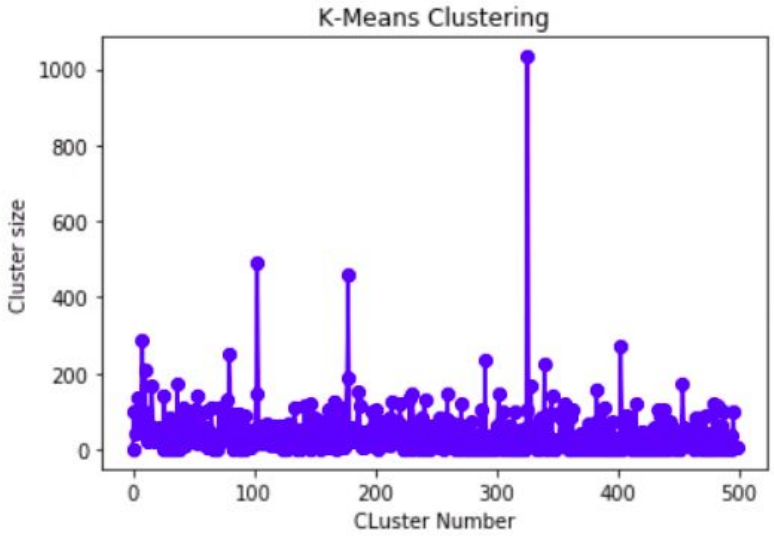
A word about metrics

- Hard to evaluate clustering algorithms when true labels are not available. We use the following metrics.
- **The Calinski-Harabasz Index (CH score)**
  - The index is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters.
  - Intuitively, the score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.
- **Silhouette Coefficient**
  - The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering.
- **Davies-Bouldin Index**
  - This index signifies the average ‘similarity’ between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.
  - Values close to zero indicate a better partitioning.

Much of the content of this slide is borrowed from: https://scikit-learn.org/stable/modules/clustering.html Last Accessed: 12/05/2019

K-Means Clustering

- **Meta**
  - Algorithm: EM based K-Means “full” algorithm. [1]
  - Cluster Centroids initialized with k-means++. [2]
  - Trained with 5 different random initializations and chosen best of 5.
  - 10 parallel threads used for data fetching.



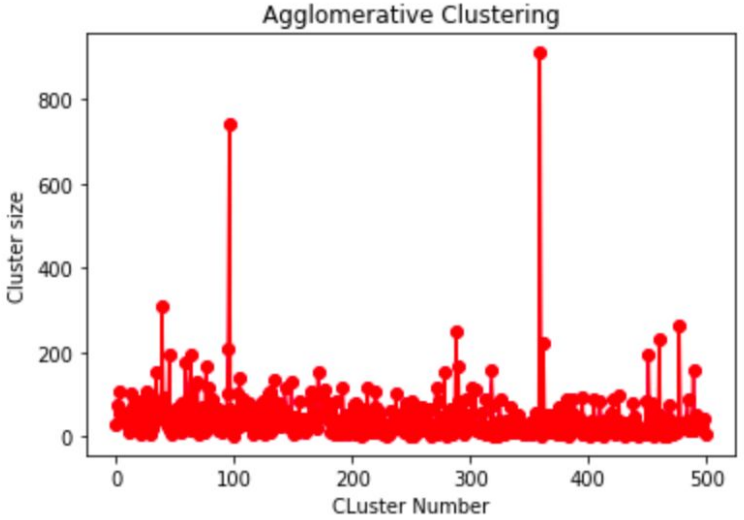
Average documents per cluster = 46.28

```
Calinski-Harasbasz Score 26.63756549021577
Davies Bouldin Score 2.9849082706410637
Silhoutte Score -0.07029791176319122
```

[1] Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982): 129-137.
[2] Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

Agglomerative Clustering

- Meta
  - Dendrogram based Hierarchical Agglomerative clustering. [1]
  - Ward based linkage with a Euclidean distance measure.
  - Constructed dendrogram to obtain 500 clusters.



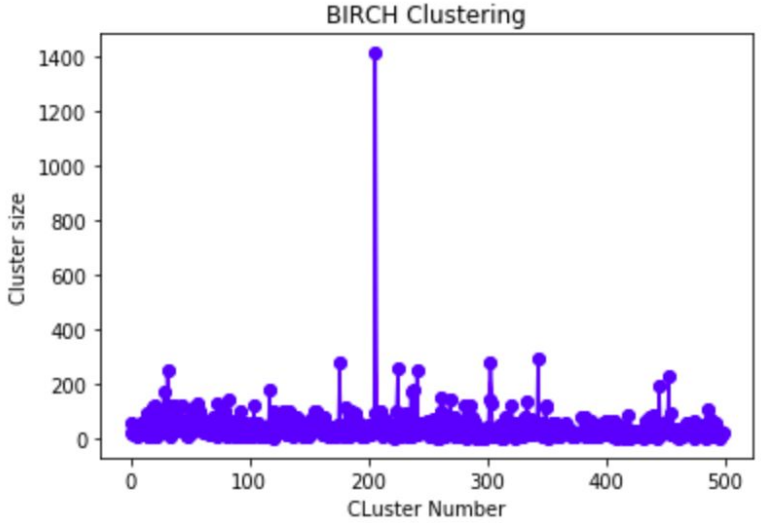
Average documents per cluster = 46.28

```
Calinski-Harasbasz Score 25.153143449371612
Davies Bouldin Score 3.4227245255504486
Silhoutte Score -0.08218313008546829
```

[1] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352.

BIRCH [1]

- **Meta**
  - Threshold: 0.5 -> Setting this value to be very low promotes splitting of clusters. [2]
  - Branching Factor: 50 -> max subclusters in a node. Additional nodes are spawned when this number is exceeded.
- **Benefits**
  - Suited for large scale databases.
  - Designed with low memory and computation footprint in mind.
  - Scales elegantly to increasing data-size. (Read easier to accommodate new incoming docs.)



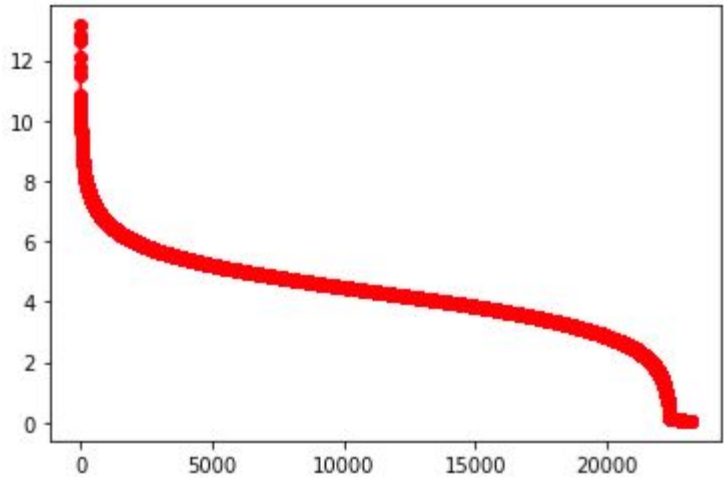
Average documents per cluster = 46.28

Calinski-Harasbasz Score 25.06948550055764
Davies Bouldin Score 3.4470936953000235
Silhoutte Score -0.07461804151535034

[1] Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." ACM Sigmod Record. Vol. 25. No. 2. ACM, 1996.
[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html

DBSCAN [1]

- **TL;DR**
  - Does not work for ETDs!
  - All documents allocated to a single cluster.
- **Benefits**
  - Designed to deal with large scale spatial databases with noise.
  - Detects and discards noisy data (Read: helpful for docs with OCR errors/garbage data).
  - Very few data specific hyper-parameters to be tuned.



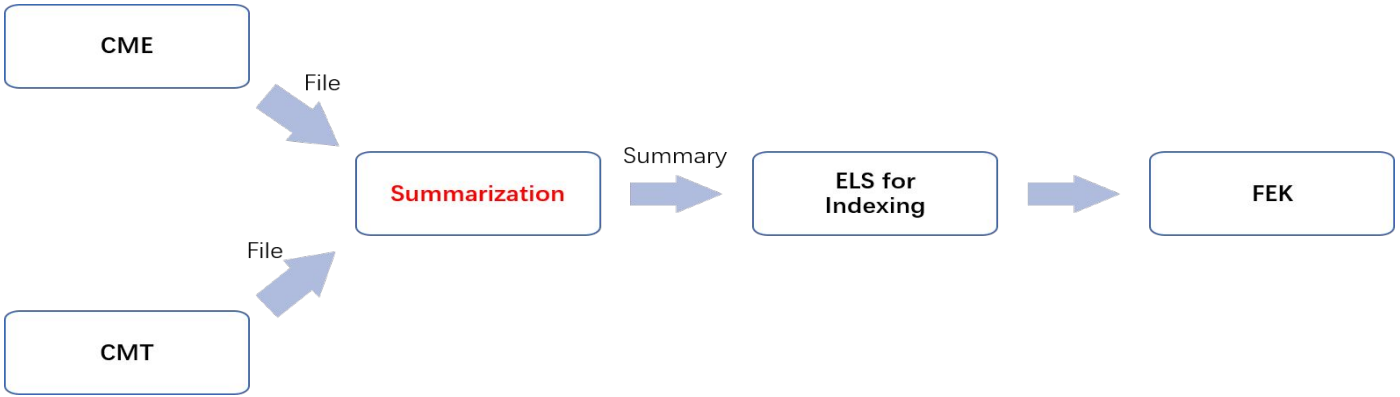
[1] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

A word about system integration

- Clustering will augment the metadata in ELS to include a cluster ID field.
- This will be used by FEK to have a “Documents similar to this” or alike button in the front-end.
- Containerized API for assigning cluster ID to new documents.

Text Summarization

- Outline for text summarization
  - Pre-processed the data for the file with too small size or too large size.
  - Built a new model for text summarization based on three different models. (Feature-based model, Graph-based model, Topic-based model)
  - Provided real summaries based on the above model for tobacco 1 million dataset.
  - Did text summarization on 20 sample dataset provided from CME team.



Why pre-processing

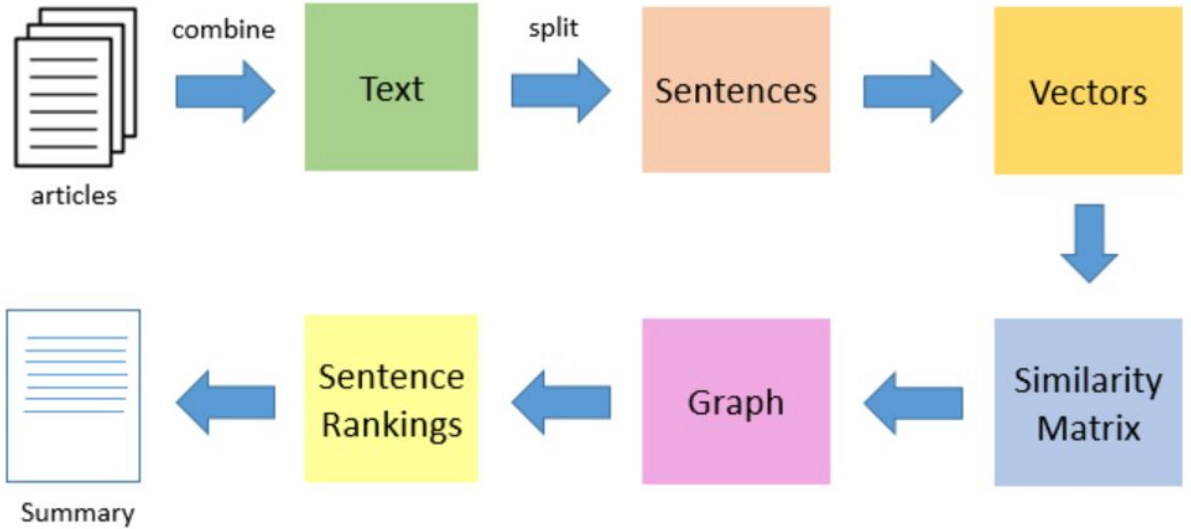
- **Eliminate Noise**
  - There are some garbage characters which may influence the result for text summarization. So we eliminate these garbage characters. (\r, \n, ~, ...)
- **Improve Efficiency**
  - For text file which includes less than 4 sentences, we don't do any summarization and just copy the original file as the summary.
  - For text file which is larger than 1Mb, we cut the whole file and do summarization separately. Otherwise, it might cause memory allocation errors.

Three Different Models

- **Feature-based Model**
    - The feature-based model will extract the features of the sentence, then a summary will be provided based on the evaluated importance. We use Luhn's algorithm which is based on TF-IDF and assigns high weights to the sentence near the beginning of the documents.

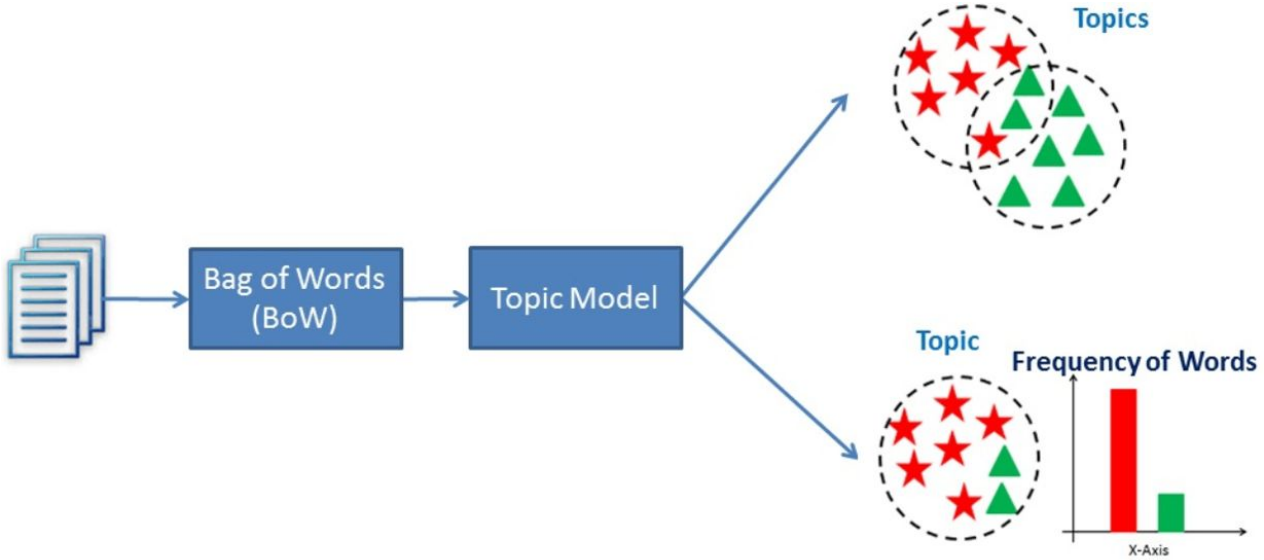
Three Different Models

- **Graph-based Model**
  - The graph-based model makes the graph from the document, then summarizes it by considering the relation between the nodes. We use TextRank, an unsupervised text summarization technique, to do summarization.



Three Different Models

- **Topic-based Model**
  - The topic-based model calculates the topic of the document and evaluates each sentence by the included topics. We use Latent Semantic Analysis, which can extract hidden semantic structures of words and sentences to detect topics.



Example

Feature-based Model

Shook, Hardy and Bacon
The industry now has
That proposal includes the
We understand that there
During the interim Dr.
The facilities proposed at
Clinical affiliations and facilities
Construction and renovation funds
In our best judgement
Should the decision be

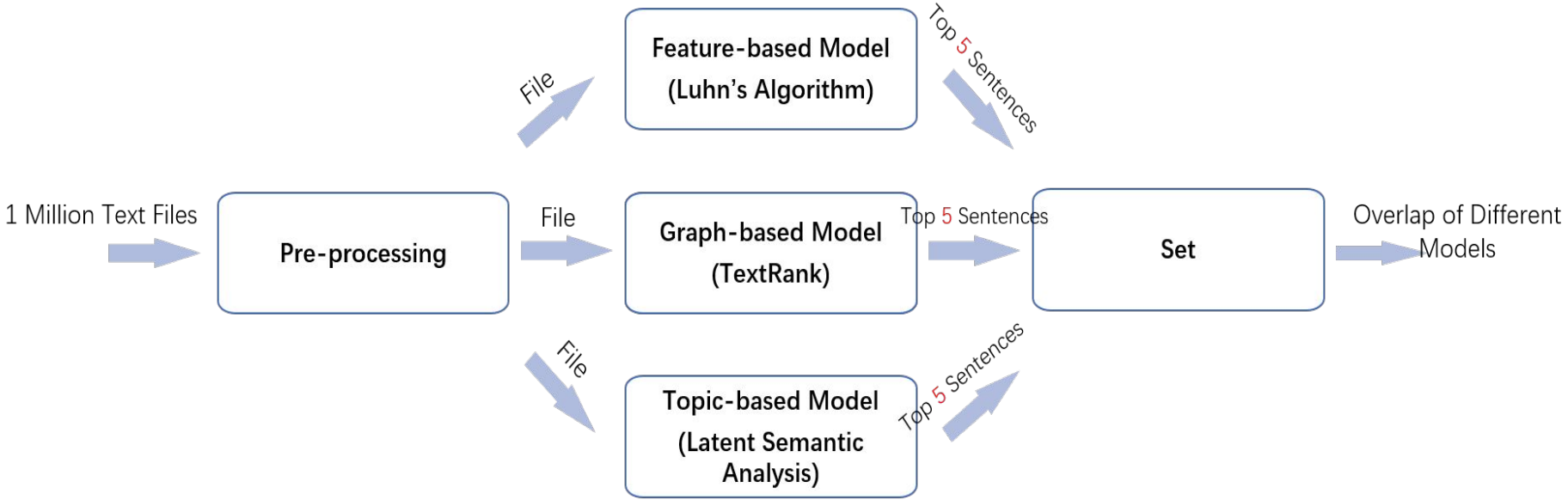
Graph-based Model

During the interim Dr
The industry now has
Clinical affiliations and facilities
Two methods of financing
The facilities proposed at
The proposal before you
That proposal includes the
We had hoped that the
In our best judgement
In closing, let me

Topic-based Model

Shook, Hardy and Bacon
We had hoped that
Although no decision was
The industry now has
That proposal includes the
We estimate that such
During the interim Dr.
In our best judgement
In closing, let me
The proposal before you

New Model (Old Version)



Number in red is the parameter which can be easily changed.

Example

Feature-based Model

Breed Sacramento Bee: J
What the really issue
And what we're looking
Dr. ERICKSON: What the
The House is considering
Sharp was one of
After a burst of
Philip Morris USA had
As often occurs with
Plan would force small

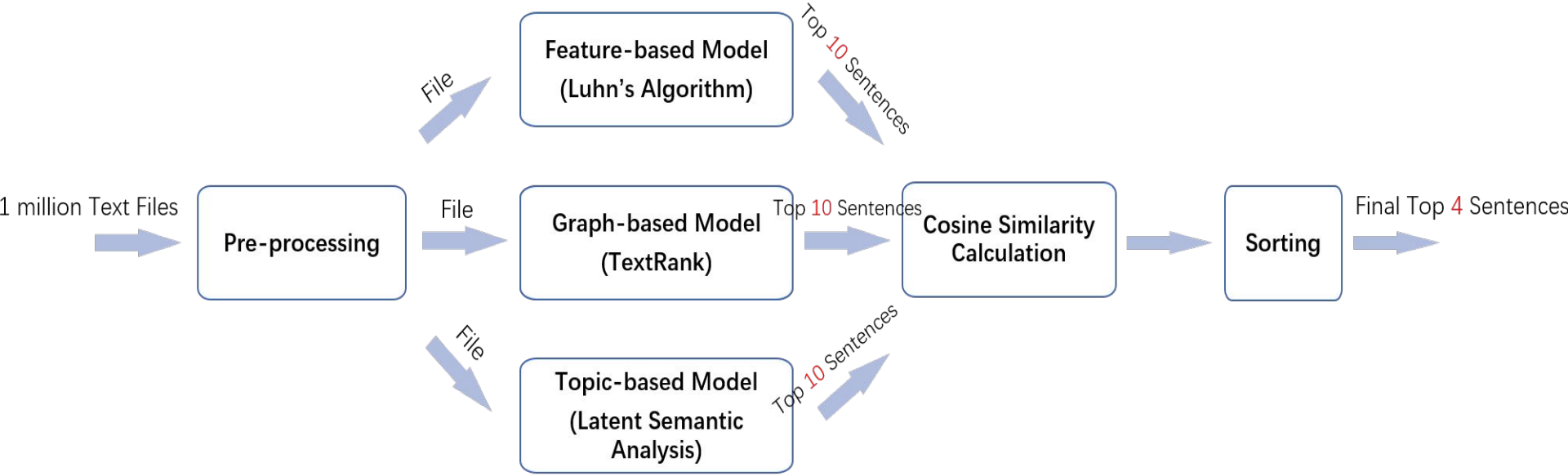
Graph-based Model

Plan would force small
It "would grant FDA
So the current effort
In exchange for the
The tobacco buyout measure
As often occurs with
But Ballin, now a
The most appealing aspect
The thinking being that
Among the key obstacles

Topic-based Model

David; Fisher, Scott; Gimbel
He says regulation of
Relatively quiet until now
Burr's proposal certainly would
Staff Photos by John
Document RNOB000020031018
Document XSN W000020031018
Philip Morris USA had
While Gregg and Kennedy
Most lawmakers agree it

New Model



Number in red is the parameter which can be easily changed.

Named-Entity Recognition (NER)

NER is about locating and classifying named entities in texts in order to recognize places, people, dates, values, organizations, etc.

- Explored different NER packages :

1. Stanford NER [1]
2. NLTK NE_Chunk [2]
3. spaCy [3]
4. Blackstone [4]
5. scispaCy [5]
6. Graphbrain [6]

References:

[1] Stanford Named Entity Recognition (NER) and Information Extraction (IE) https://nlp.stanford.edu/ner/
[2] nltk.chunk package, NLTK 3.4.5 documentation https://www.nltk.org/api/nltk.chunk.html
[3] spaCy: Industrial-Strength Natural Language Processing https://spacy.io/
[4] Blackstone: Model for NLP on unstructured legal text https://spacy.io/universe/project/blackstone
[5] scispaCy: Models for scientific/biomedical documents https://spacy.io/universe/project/scispacy
[6] Graphbrain: Automated meaning extraction and text understanding https://spacy.io/universe/project/graphbrain

Named-Entity Recognition (NER)

NER is about locating and classifying named entities in texts in order to recognize places, people, dates, values, organizations, etc.

- Explored different NER packages :

1. Stanford NER [1]
2. NLTK NE_Chunk [2]
3. **spaCy [3]**

4. Blackstone [4]
5. scispaCy [5]
6. Graphbrain [6]

References:

[1] Stanford Named Entity Recognition (NER) and Information Extraction (IE) https://nlp.stanford.edu/ner/
[2] nltk.chunk package, NLTK 3.4.5 documentation https://www.nltk.org/api/nltk.chunk.html
[3] spaCy: Industrial-Strength Natural Language Processing https://spacy.io/
[4] Blackstone: Model for NLP on unstructured legal text https://spacy.io/universe/project/blackstone
[5] scispaCy: Models for scientific/biomedical documents https://spacy.io/universe/project/scispacy
[6] Graphbrain: Automated meaning extraction and text understanding https://spacy.io/universe/project/graphbrain

Named-Entity Recognition (NER)

- spaCy provided the best results
- spaCy is used for Named-Entity Recognition on the entire Tobacco dataset.

spaCy

Example

Original Sentence:

The witness, senior vice-president and controller at R. J. Reynolds Tobacco Holding Inc., was deposed by the plaintiffs. He described the financial status of the holding company and its economic activities. He indicated that industry changes, corporate changes, market changes, structural changes, and some legal developments have all had an adverse effect on the profitability of the company. The witness also noted that advertising and promotion restrictions placed on them in 1998 by the Master Settlement Agreement had caused a drop in sales volume. He said that punitive damage awards would have a devastating effect on the company, although he declined to say whether bankruptcy was being considered.

Extracted Entities

Type: ORG, Value: R. J. Reynolds Tobacco Holding Inc.
Type: DATE, Value: 1998
Type: LAW, Value: the Master Settlement Agreement

spaCy

Example

Original Sentence:

The witness, senior vice-president and controller at **R. J. Reynolds Tobacco Holding Inc.**, was deposed by the plaintiffs. He described the financial status of the holding company and its economic activities. He indicated that industry changes, corporate changes, market changes, structural changes, and some legal developments have all had an adverse effect on the profitability of the company. The witness also noted that advertising and promotion restrictions placed on them in **1998** by the **Master Settlement Agreement** had caused a drop in sales volume. He said that punitive damage awards would have a devastating effect on the company, although he declined to say whether bankruptcy was being considered.

Extracted Entities

Type: ORG, Value: R. J. Reynolds Tobacco Holding Inc. ✅
Type: DATE, Value: 1998 ✅
Type: LAW, Value: the Master Settlement Agreement ✅

spaCy

Example

Original Sentence:

The witness, Director of Marketing Research at Philip Morris, was deposed by the plaintiffs. He reviewed his previous depositions and trail testimony, as well as the contract work that he has done for Philip Morris. He explained that the contract work consisted of showing advertising or packaging and obtaining information on consumer reactions. He reviewed the organizational structure of the Marketing and Research department of Philip Morris. The witness listed the various companies from which Philip Morris obtained consumer information. He maintained that Philip Morris only conducted studies on people over the age of 18. He explained the importance of having highly reliable information about legal age smokers in order to accurately project future industry sales and brand sales. He described Philip Morris' use of publicly available information and studies on smoking behavior. He commented on surveys in which adults were asked about their age of smoking initiation.; Roper

Extracted Entities

Type: ORG, Value: Marketing Research
Type: ORG, Value: Philip Morris
Type: ORG, Value: Philip Morris
Type: ORG, Value: the Marketing and Research
Type: ORG, Value: Philip Morris
Type: ORG, Value: Philip Morris
Type: ORG, Value: Philip Morris
Type: DATE, Value: the age of 18
Type: ORG, Value: Philip Morris'
Type: PERSON, Value: Roper

spaCy

Example

Original Sentence:

The witness, Director of **Marketing Research** at **Philip Morris**, was deposed by the plaintiffs. He reviewed his previous depositions and trail testimony, as well as the contract work that he has done for **Philip Morris**. He explained that the contract work consisted of showing advertising or packaging and obtaining information on consumer reactions. He reviewed the organizational structure of **the Marketing and Research** department of **Philip Morris**. The witness listed the various companies from which **Philip Morris** obtained consumer information. He maintained that **Philip Morris** only conducted studies on people over **the age of 18**. He explained the importance of having highly reliable information about legal age smokers in order to accurately project future industry sales and brand sales. He described **Philip Morris'** use of publicly available information and studies on smoking behavior. He commented on surveys in which adults were asked about their age of smoking initiation.; **Roper**

Extracted Entities

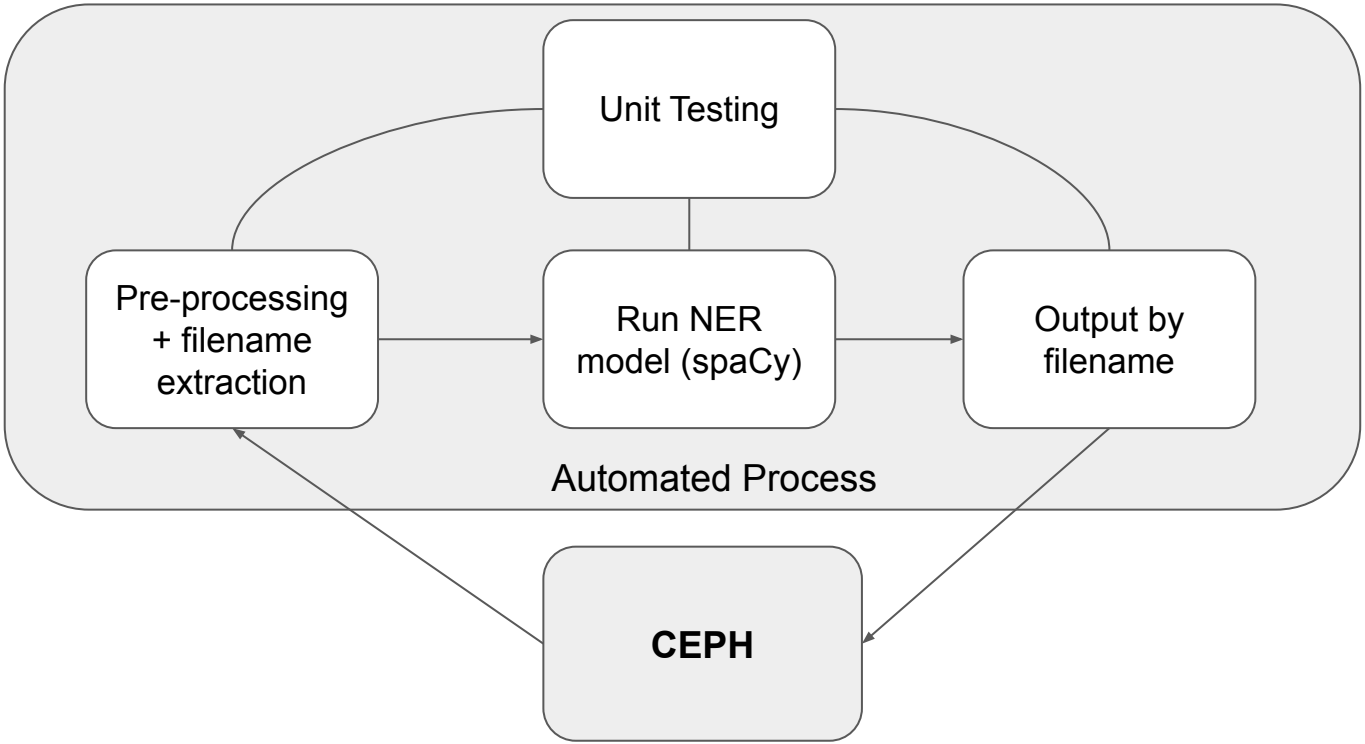
Type: ORG, Value: Marketing Research ✅
Type: ORG, Value: Philip Morris ✅
Type: ORG, Value: Philip Morris ✅
Type: ORG, Value: the Marketing and Research ✅
Type: ORG, Value: Philip Morris ✅

Type: ORG, Value: Philip Morris ✅
Type: ORG, Value: Philip Morris ✅
Type: DATE, Value: the age of 18 ✅
Type: ORG, Value: Philip Morris' ✅
Type: PERSON, Value: Roper ✅

spaCy Models

1.	en_core_web_sm		11 MB
2.	en_core_web_md		91 MB
3.	en_core_web_lg		789 MB
4.	en_trf_bertbaseuncased_lg	(Google & Hugging Face)	387 MB
5.	en_trf_robertabase_lg	(Facebook & Hugging Face)	278 MB
6.	en_trf_distilbertbaseuncased_lg	(Hugging Face)	233 MB
7.	en_trf_xlnetbasecased_lg	(CMU & Google Brain)	413 MB

NER System Architecture



NER Automation

- Scripts for automation of NER on tobacco dataset
- Automation has been performed on a sample dataset on local machine.
- Results of NER is stored in a text file for ingestion by ELS team.
- Kev value pairs of NER

Example: Document ID: jtvf0005

Attendance at PR meeting September 10, 1958; James P. Richards Robert K. He~jnann 0. D. Campbell 6ene L. Cooper W. S. Cutchins Margaret Carson H. C. Robinson Jr. Dan Provost James C. Bowling Jokm Scott Fones Rex Lardner John Jones Richard W. Darrow Carl C. Thompson Leonard S. Zatm Kenneth L. Austin D7alco'-m Jo'nnson W. T. Hoyt The Tobacco Institute, Inc, The Amer:can Tobacco Company Braun and Company, Inc. Braun and Company, Inc. Brown and Williamson Tobacco Corp. M. Carson, Inc. Liggett & Myers Tobacco Company D:cCann-Erickson, Inc. Philip Morris, Inc. Benjamin Sonnenberg Sidney J. Wain, Inc. Sidney J. Wain, Inc. Hill end Hnowlton, Inc. F.111 and Knowlton, Inc. Hill and Knowlton, Inc. Hil'_ and Knowlton, Inc. Hi1= and Kaowlton, Inc. Tobacco Industry Research Conmittee

Example: Document ID: jtvf0005 - NER results

- **jtvf0005.txt:**

[('DATE', September 10, 1958), ('PERSON', James P. Richards), ('PERSON', Robert K.), ('NORP', D.), ('ORG', Campbell), ('PERSON', James C. Bowling), ('PERSON', John Jones Richard W. Darrow), ('PERSON', Carl C. Thompson), ('ORG', The Tobacco Institute), ('PERSON', Inc), ('ORG', The Amer:can Tobacco Company Braun and Company), ('PERSON', M. Carson), ('ORG', Liggett & Myers Tobacco Company D), ('ORG', cCann-Erickson), ('ORG', Philip Morris), ('PERSON', Benjamin Sonnenberg Sidney J. Wain), ('PERSON', Sidney J. Wain), ('ORG', Inc. Hill end), ('GPE', Hnowlton), ('ORG', Inc. F.111), ('GPE', Knowlton), ('ORG', Inc. Hill), ('GPE', Knowlton), ('ORG', Inc. Hil'), ('WORK_OF_ART', _ and Knowlton, Inc. Hi1=), ('GPE', Kaowlton)]

Sentiment Analysis

- Flair [1]
- Twitter Emotion Recognition [2]
- Empath [3]
- SenticNet [4]

[1] Flair: Pooled Contextualized Embeddings for Named Entity Recognition, https://github.com/zalandoresearch/flair
[2] Twitter Emotion Recognition, https://github.com/nikicc/twitter-emotion-recognition
[3] Empath: Understanding Topic Signals in Large-Scale Text, https://hci.stanford.edu/publications/2016/ethan/empath-chi-2016.pdf
[4] SenticNet: Emotion Recognition in Conversations, https://github.com/SenticNet/conv-emotion

Sentiment Analysis

- Flair [1]
- Twitter Emotion Recognition [2]
- **Empath [3]**
- SenticNet [4]

[1] Flair: Pooled Contextualized Embeddings for Named Entity Recognition, https://github.com/zalandoresearch/flair
[2] Twitter Emotion Recognition, https://github.com/nikicc/twitter-emotion-recognition
[3] Empath: Understanding Topic Signals in Large-Scale Text, https://hci.stanford.edu/publications/2016/ethan/empath-chi-2016.pdf
[4] SenticNet: Emotion Recognition in Conversations, https://github.com/SenticNet/conv-emotion

Empath: Categories

social media	war	violence	technology	fear	pain	hipster	contempt
facebook	attack	hurt	ipad	horror	hurt	vintage	disdain
instagram	battlefield	break	internet	paralyze	pounding	trendy	mockery
notification	soldier	bleed	download	dread	sobbing	fashion	grudging
selfie	troop	broken	wireless	scared	gasp	designer	haughty
account	army	scar	computer	tremor	torment	artsy	caustic
timeline	enemy	hurting	email	despair	groan	1950s	censure
follower	civilian	injury	virus	panic	stung	edgy	sneer

Table 1. Empath can analyze text across hundreds of data-driven categories. Here we provide a sample of representative terms in 8 sample categories.

Total categories: 194
Total models: 3

Empath: Lexical Categorization

[('achievement', 0.0), ('affection', 0.0020366598778004073), ('aggression', 0.0020366598778004073), ('air_travel', 0.0), ('alcohol', 0.0), ('ancient', 0.0), ('anger', 0.0), ('animal', 0.0), ('anonymity', 0.0), ('anticipation', 0.0), ('appearance', 0.0), ('art', 0.012219959266802444), ('attractive', 0.0), ('banking', 0.0), ('beach', 0.0), ('beauty', 0.0), ('blue_collar_job', 0.0), ('body', 0.0), ('breaking', 0.0), ('business', 0.0), ('car', 0.0), ('celebration', 0.0), ('cheerfulness', 0.0), ('childish', 0.0), ('children', 0.0), ('cleaning', 0.0), ('clothing', 0.0), ('cold', 0.0), ('college', 0.006109979633401222), ('communication', 0.008146639511201629), ('competing', 0.0), ('computer', 0.006109979633401222), ('confusion', 0.0), ('contentment', 0.0020366598778004073), ('cooking', 0.0), ('crime', 0.0), ('dance', 0.0020366598778004073), ('death', 0.0), ('deception', 0.0), ('disappointment', 0.0), ('disgust', 0.0), ('dispute', 0.004073319755600814), ('divine', 0.0), ('domestic_work', 0.0), ('dominant_heirarchical', 0.0), ('dominant_personality', 0.0), ('driving', 0.0), ('eating', 0.0), ('economics', 0.0020366598778004073), ('emotional', 0.0), ('envy', 0.0), ('exasperation', 0.0), ('exercise', 0.004073319755600814), ('exotic', 0.0), ('fabric', 0.0), ('family', 0.0), ('farming', 0.0), ('fashion', 0.0), ('fear', 0.004073319755600814), ('feminine', 0.0), ('fight', 0.006109979633401222), ('fire', 0.0), ('friends', 0.0020366598778004073), ('fun', 0.0), ('furniture', 0.0), ('gain', 0.0020366598778004073), ('giving', 0.0), ('government', 0.0), ('hate', 0.0020366598778004073), ('healing', 0.006109979633401222), ('health', 0.0), ('hearing', 0.0), ('help', 0.004073319755600814), ('heroic', 0.004073319755600814), ('hiking', 0.0), ('hipster', 0.006109979633401222), ('home', 0.0), ('horror', 0.0), ('hygiene', 0.0), ('independence', 0.0020366598778004073), ('injury', 0.0), ('internet', 0.006109979633401222), ('irritability', 0.0), ('journalism', 0.0020366598778004073), ('joy', 0.0020366598778004073), ('kill', 0.0), ('law', 0.0), ('leader', 0.0), ('legend', 0.0), ('leisure', 0.0), ('liquid', 0.0), ('listen', 0.0020366598778004073), ('love', 0.0020366598778004073), ('lust', 0.0020366598778004073), ('magic', 0.0), ('masculine', 0.0), ('medical_emergency', 0.0), ('medieval', 0.0), ('meeting', 0.018329938900203666), ('messaging', 0.0), ('military', 0.004073319755600814), ('money', 0.0020366598778004073), ('monster', 0.0), ('morning', 0.0), ('movement', 0.008146639511201629), ('music', 0.0020366598778004073), ('musical', 0.0020366598778004073), ('negative_emotion', 0.0), ('neglect', 0.0), ('negotiate', 0.0), ('nervousness', 0.0020366598778004073), ('night', 0.0), ('noise', 0.0), ('occupation', 0.0), ('ocean', 0.0), ('office', 0.0), ('optimism', 0.008146639511201629), ('order', 0.0), ('pain', 0.0020366598778004073), ('party', 0.0), ('payment', 0.0020366598778004073), ('pet', 0.0), ('philosophy', 0.004073319755600814), ('phone', 0.0), ('plant', 0.0), ('play', 0.0), ('politeness', 0.0020366598778004073), ('politics', 0.0), ('poor', 0.0), ('positive_emotion', 0.010183299389002037), ('power', 0.004073319755600814), ('pride', 0.0), ('prison', 0.0), ('programming', 0.006109979633401222), ('rage', 0.0), ('reading', 0.008146639511201629), ('real_estate', 0.0), ('religion', 0.0), ('restaurant', 0.0), ('ridicule', 0.0), ('royalty', 0.0), ('rural', 0.0), ('sadness', 0.0020366598778004073), ('sailing', 0.0), ('school', 0.014256619144602852), ('science', 0.006109979633401222), ('sexual', 0.0), ('shame', 0.0020366598778004073), ('shape_and_size', 0.0020366598778004073), ('ship', 0.0), ('shopping', 0.0), ('sleep', 0.0), ('smell', 0.0), ('social_media', 0.016293279022403257), ('sound', 0.0), ('speaking', 0.008146639511201629), ('sports', 0.0), ('stealing', 0.0), ('strength', 0.0020366598778004073), ('suffering', 0.0020366598778004073), ('superhero', 0.0), ('surprise', 0.0), ('swearing_terms', 0.0), ('swimming', 0.0), ('sympathy', 0.004073319755600814), ('technology', 0.006109979633401222), ('terrorism', 0.0), ('timidity', 0.0), ('tool', 0.0), ('torment', 0.0), ('tourism', 0.0), ('toy', 0.0), ('traveling', 0.0), ('trust', 0.0020366598778004073), ('ugliness', 0.0), ('urban', 0.0), ('vacation', 0.0), ('valuable', 0.0020366598778004073), ('vehicle', 0.0), ('violence', 0.0), ('war', 0.0), ('warmth', 0.0), ('water', 0.0), ('weakness', 0.0), ('wealthy', 0.004073319755600814), ('weapon', 0.0), ('weather', 0.0), ('wedding', 0.0), ('white_collar_job', 0.0), ('work', 0.0), ('worship', 0.0), ('writing', 0.0020366598778004073), ('youth', 0.0), ('zest', 0.0020366598778004073)]

Empath: Lexical Categorization

[('meeting', 0.018329938900203666), ('social_media', 0.016293279022403257), ('school', 0.014256619144602852), ('art', 0.012219959266802444), ('positive_emotion', 0.010183299389002037), ('optimism', 0.008146639511201629), ('reading', 0.008146639511201629), ('movement', 0.008146639511201629), ('communication', 0.008146639511201629), ('speaking', 0.008146639511201629), ('computer', 0.006109979633401222), ('college', 0.006109979633401222), ('hipster', 0.006109979633401222), ('internet', 0.006109979633401222), ('healing', 0.006109979633401222), ('programming', 0.006109979633401222), ('fight', 0.006109979633401222), ('science', 0.006109979633401222), ('technology', 0.006109979633401222), ('help', 0.004073319755600814), ('dispute', 0.004073319755600814), ('wealthy', 0.004073319755600814), ('exercise', 0.004073319755600814), ('fear', 0.004073319755600814), ('heroic', 0.004073319755600814), ('military', 0.004073319755600814), ('sympathy', 0.004073319755600814), ('power', 0.004073319755600814), ('philosophy', 0.004073319755600814), ('dance', 0.002036659877800407), ('money', 0.002036659877800407), ('hate', 0.002036659877800407), ('aggression', 0.002036659877800407), ('nervousness', 0.002036659877800407), ('suffering', 0.002036659877800407), ('journalism', 0.002036659877800407), ('independence', 0.002036659877800407), ('zest', 0.002036659877800407), ('love', 0.002036659877800407), ('trust', 0.002036659877800407), ('music', 0.002036659877800407), ('politeness', 0.002036659877800407), ('listen', 0.002036659877800407), ('gain', 0.002036659877800407), ('valuable', 0.002036659877800407), ('sadness', 0.002036659877800407), ('joy', 0.002036659877800407), ('affection', 0.002036659877800407), ('lust', 0.002036659877800407), ('shame', 0.002036659877800407), ('economics', 0.002036659877800407), ('strength', 0.002036659877800407), ('shape_and_size', 0.002036659877800407), ('pain', 0.002036659877800407), ('friends', 0.002036659877800407), ('payment', 0.002036659877800407), ('contentment', 0.002036659877800407), ('writing', 0.002036659877800407), ('musical', 0.002036659877800407), ('office', 0.0), ('wedding', 0.0), ('domestic_work', 0.0), ('sleep', 0.0), ('medical_emergency', 0.0), ('cold', 0.0), ('cheerfulness', 0.0), ('occupation', 0.0), ('envy', 0.0), ('anticipation', 0.0), ('family', 0.0), ('vacation', 0.0), ('crime', 0.0), ('attractive', 0.0), ('masculine', 0.0), ('prison', 0.0), ('health', 0.0), ('pride', 0.0), ('government', 0.0), ('weakness', 0.0), ('horror', 0.0), ('swearing_terms', 0.0), ('leisure', 0.0), ('royalty', 0.0), ('tourism', 0.0), ('furniture', 0.0), ('magic', 0.0), ('beach', 0.0), ('morning', 0.0), ('banking', 0.0), ('night', 0.0), ('kill', 0.0), ('blue_collar_job', 0.0), ('ridicule', 0.0), ('play', 0.0), ('stealing', 0.0), ('real_estate', 0.0), ('home', 0.0), ('divine', 0.0), ('sexual', 0.0), ('irritability', 0.0), ('superhero', 0.0), ('business', 0.0), ('driving', 0.0), ('pet', 0.0), ('childish', 0.0), ('cooking', 0.0), ('exasperation', 0.0), ('religion', 0.0), ('surprise', 0.0), ('worship', 0.0), ('leader', 0.0), ('body', 0.0), ('noise', 0.0), ('eating', 0.0), ('medieval', 0.0), ('confusion', 0.0), ('water', 0.0), ('sports', 0.0), ('death', 0.0), ('legend', 0.0), ('celebration', 0.0), ('restaurant', 0.0), ('violence', 0.0), ('dominant_heirarchical', 0.0), ('neglect', 0.0), ('swimming', 0.0), ('exotic', 0.0), ('hiking', 0.0), ('hearing', 0.0), ('order', 0.0), ('hygiene', 0.0), ('weather', 0.0), ('anonymity', 0.0), ('ancient', 0.0), ('deception', 0.0), ('fabric', 0.0), ('air_travel', 0.0), ('dominant_personality', 0.0), ('vehicle', 0.0), ('toy', 0.0), ('farming', 0.0), ('war', 0.0), ('urban', 0.0), ('shopping', 0.0), ('disgust', 0.0), ('fire', 0.0), ('tool', 0.0), ('phone', 0.0), ('sound', 0.0), ('injury', 0.0), ('sailing', 0.0), ('rage', 0.0), ('work', 0.0), ('appearance', 0.0), ('warmth', 0.0), ('youth', 0.0), ('fun', 0.0), ('emotional', 0.0), ('traveling', 0.0), ('fashion', 0.0), ('ugliness', 0.0), ('torment', 0.0), ('anger', 0.0), ('politics', 0.0), ('ship', 0.0), ('clothing', 0.0), ('car', 0.0), ('breaking', 0.0), ('white_collar_job', 0.0), ('animal', 0.0), ('party', 0.0), ('terrorism', 0.0), ('smell', 0.0), ('disappointment', 0.0), ('poor', 0.0), ('plant', 0.0), ('beauty', 0.0), ('timidity', 0.0), ('negotiate', 0.0), ('negative_emotion', 0.0), ('cleaning', 0.0), ('messaging', 0.0), ('competing', 0.0), ('law', 0.0), ('achievement', 0.0), ('alcohol', 0.0), ('liquid', 0.0), ('feminine', 0.0), ('weapon', 0.0), ('children', 0.0), ('monster', 0.0), ('ocean', 0.0), ('giving', 0.0), ('rural', 0.0)]

Empath: Lexical Categorization

[('meeting', 0.018329938900203666), ('social_media', 0.016293279022403257), ('school', 0.014256619144602852), ('art', 0.012219959266802444), ('positive_emotion', 0.010183299389002037), ('optimism', 0.008146639511201629), ('reading', 0.008146639511201629), ('movement', 0.008146639511201629), ('communication', 0.008146639511201629), ('speaking', 0.008146639511201629), ('computer', 0.006109979633401222), ('college', 0.006109979633401222), ('hipster', 0.006109979633401222), ('internet', 0.006109979633401222), ('healing', 0.006109979633401222), ('programming', 0.006109979633401222), ('fight', 0.006109979633401222), ('science', 0.006109979633401222), ('technology', 0.006109979633401222), ('help', 0.004073319755600814), ('dispute', 0.004073319755600814), ('wealthy', 0.004073319755600814), ('exercise', 0.004073319755600814), ('fear', 0.004073319755600814), ('heroic', 0.004073319755600814), ('military', 0.004073319755600814), ('sympathy', 0.004073319755600814), ('power', 0.004073319755600814), ('philosophy', 0.004073319755600814), ('dance', 0.002036659877800407), ('money', 0.002036659877800407), ('hate', 0.002036659877800407), ('aggression', 0.002036659877800407), ('nervousness', 0.002036659877800407), ('suffering', 0.002036659877800407), ('journalism', 0.002036659877800407), ('independence', 0.002036659877800407), ('zest', 0.002036659877800407), ('love', 0.002036659877800407), ('trust', 0.002036659877800407), ('music', 0.002036659877800407), ('politeness', 0.002036659877800407), ('listen', 0.002036659877800407), ('gain', 0.002036659877800407), ('valuable', 0.002036659877800407), ('sadness', 0.002036659877800407), ('joy', 0.002036659877800407), ('affection', 0.002036659877800407), ('lust', 0.002036659877800407), ('shame', 0.002036659877800407), ('economics', 0.002036659877800407), ('strength', 0.002036659877800407), ('shape_and_size', 0.002036659877800407), ('pain', 0.002036659877800407), ('friends', 0.002036659877800407), ('payment', 0.002036659877800407), ('contentment', 0.002036659877800407), ('writing', 0.002036659877800407), ('musical', 0.002036659877800407)]

How many categories to consider?

Total categories: 194
Total models: 3

Basic Emotions

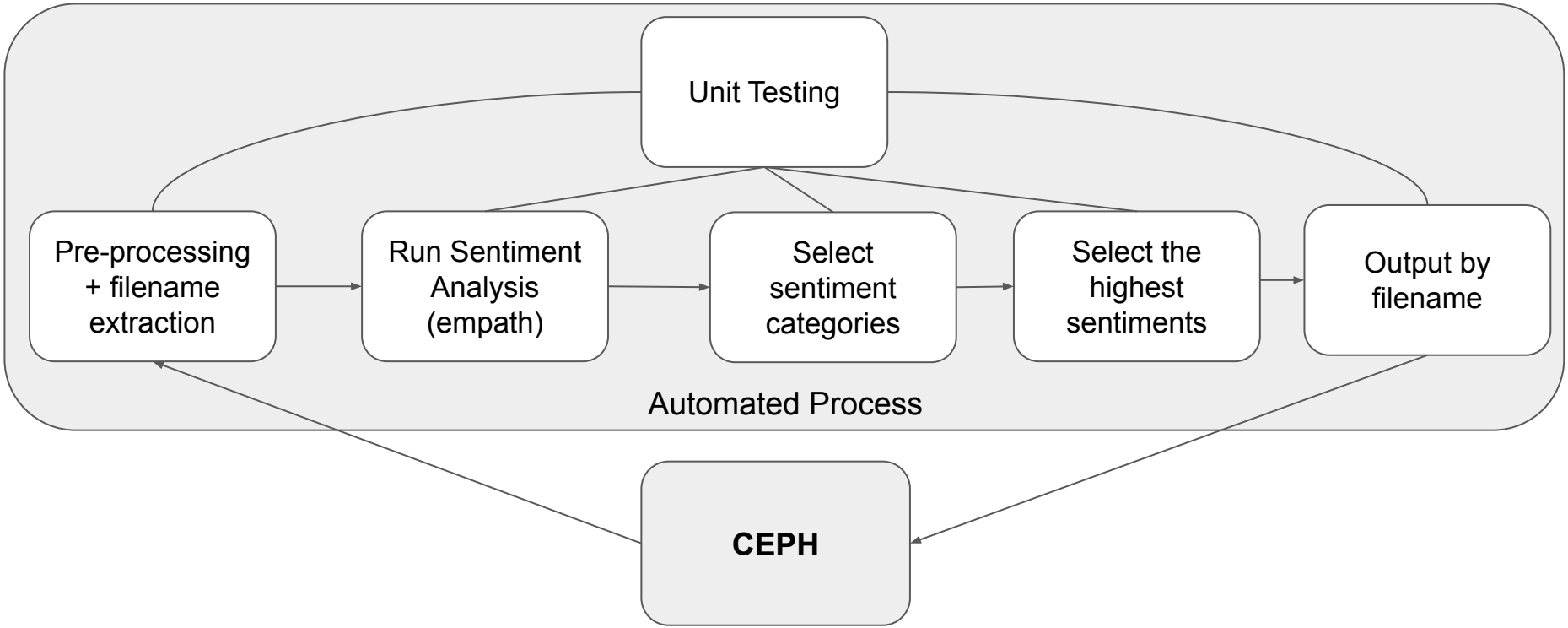
- Ekman's six basic emotions [1]
- Plutchik's eight basic emotions [2]
- Profile of Mood States (POMS) six mood states [3]

References:
[1] Ekman, Paul. "Basic emotions." Handbook of cognition and emotion 98.45-60 (1999): 16.
[2] Plutchik, Robert. "Emotions: A general psychoevolutionary theory." Approaches to emotion 1984 (1984): 197-219.
[3] Curran, Shelly L., Michael A. Andrykowski, and Jamie L. Studts. "Short form of the profile of mood states (POMS-SF): psychometric information." Psychological assessment 7.1 (1995): 80.

Six basic emotions

- Love
- Hate
- Joy
- Fear
- Surprise
- Envy

Sentiment Analysis System Architecture



Recommender System

- System that is capable of predicting the future preference of a set of items for a user, and recommend the top items.

Implementation details :

- Identified a sample dataset of user logs
- Implemented content based and collaborative filtering recommendation techniques on this sample dataset

Why this sample dataset ?

- Our entire search engine was not completely integrated at that time and we wanted to show a prototype of implementation
- The selected dataset is based on real logs and has fields similar to our search logs

SAMPLE DATASET: CI&T's Internal Communication platform (DeskDrop)[1]

- Contains a real sample of 12 months logs (Mar. 2016 - Feb. 2017) and 73k logged users interactions
- 1140 total users, 2926 documents
- Has fields such as Person ID, Content ID, Session ID, Timestamp, etc

[1] https://www.kaggle.com/gspmoreira/articles-sharing-reading-from-cit-deskdrop

Content based recommendation

- Recommends items that are similar to those that a user liked in the past

STEPS:

1) Build user profile by constructing item profile for all the items the user has interacted with using TF-IDF.
2) Get items which are similar to the user profile - Cosine Similarity between user profile and TF-IDF Matrix
3) Sort the similar items and recommend items to the user.

Evaluation result of Content based filtering :

- **Recall@5 = 0.4145**
- **Recall@10 = 0.5241**

Collaborative Filtering Model

- **User-based Approach:** Uses memory of previous users interactions to compute similarity based on items they have interacted with.
- **Item-based Approach**: Compute item similarities based on users that have interacted with them.

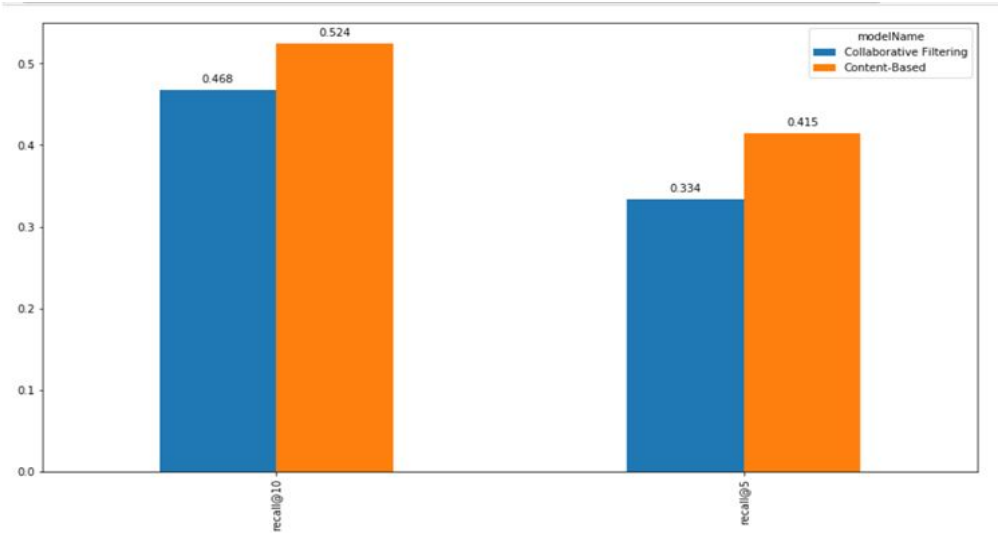
Matrix Factorization:

- User-item matrix is compressed to low-dimensional representation in terms of latent factors.
- SVD (Singular value decomposition) Latent factor model is used.

Evaluation Result of Collaborative Filtering

- **Recall@5 =** 33.4%
- **Recall@10 =** 46.81%

Performance comparison:



- **Content based** : Clustering ✅
- **Collaborative filtering** : User logs

User Logs: What we have currently

FEK Logs :


```
{
        "status": 200,
        "message": "Success",
        "data": {
                "user": {
                        "username": "Eddy",
                        "email": "no email given"
                },
                "activity": {
                        "url":
"http://localhost:9200/etd_metadata/_msearch?",
                        "search_text": "title-none: Immersion",
                        "filters": {}
                },
                "dataset": "etd",
                "time": "2019-11-08 19:26:49.530631",
                "ip": "127.0.0.1"
        }
}
```


ELS Logs:


```
{  "type": "index_search_slowlog",
        "timestamp": "2019-12-04T01:09:09,002Z",
        "level": "WARN",
        "component": "i.s.s.query",
        "cluster.name": "elasticsearch",
        "node.name": "elasticsearch-master-0",
        "message": "[etd_metadata][0]",
        "took": "930.9ms",
        "took_millis": "930",
        "total_hits": "19 hits",
        "search_type": "QUERY_THEN_FETCH",
        "total_shards": "1",
        "source": "{\"query\"
:{\"term\":{\"title-none\":{\"value\":\"data\",\"boost\":1.0}}}}",
        "cluster.uuid": "M7gJSQVkSYi3THDYCTvIew",
        "node.id": "nXkX9qONS2y0g5WB8NGezQ"}
```

Comparison between log fields

User logs from sample datasets:

Event Type	Content ID	Person ID	Session ID
{ 'View' : 1.0, 'Like' : 2.0, 'Bookmark':2.5, 'Follow': 3.0, 'Comment Created': 4.0}			

User Log Fields obtained from FEK and ELS logs:

User ID	Search Query

Missing Field:

Document ID/ List of Documents viewed per query

Recommender System: Future scope

- Add ‘document viewed’ field in FEK/ELS metadata
- Collect significant number of user logs
- Create user item matrix
- Recommend items tailored to the user’s preference

Thank you