Tobacco Testimonies

Team Members: Nick Onofrio, Nick Sorkin, Campbell Johnson, Michael DiFrancisco, and Devin Venetsanos

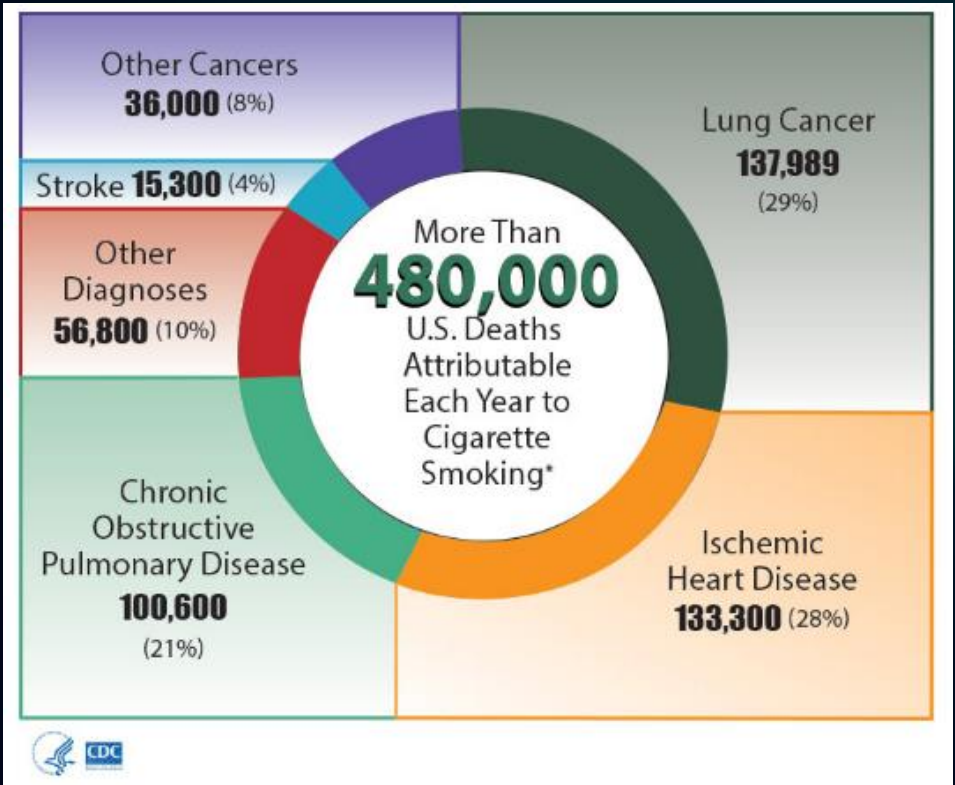
Multimedia, Hypertext, and Information Accessx - Professor Fox

Virginia Tech, Blacksburg VA 24061, 2/14/2019

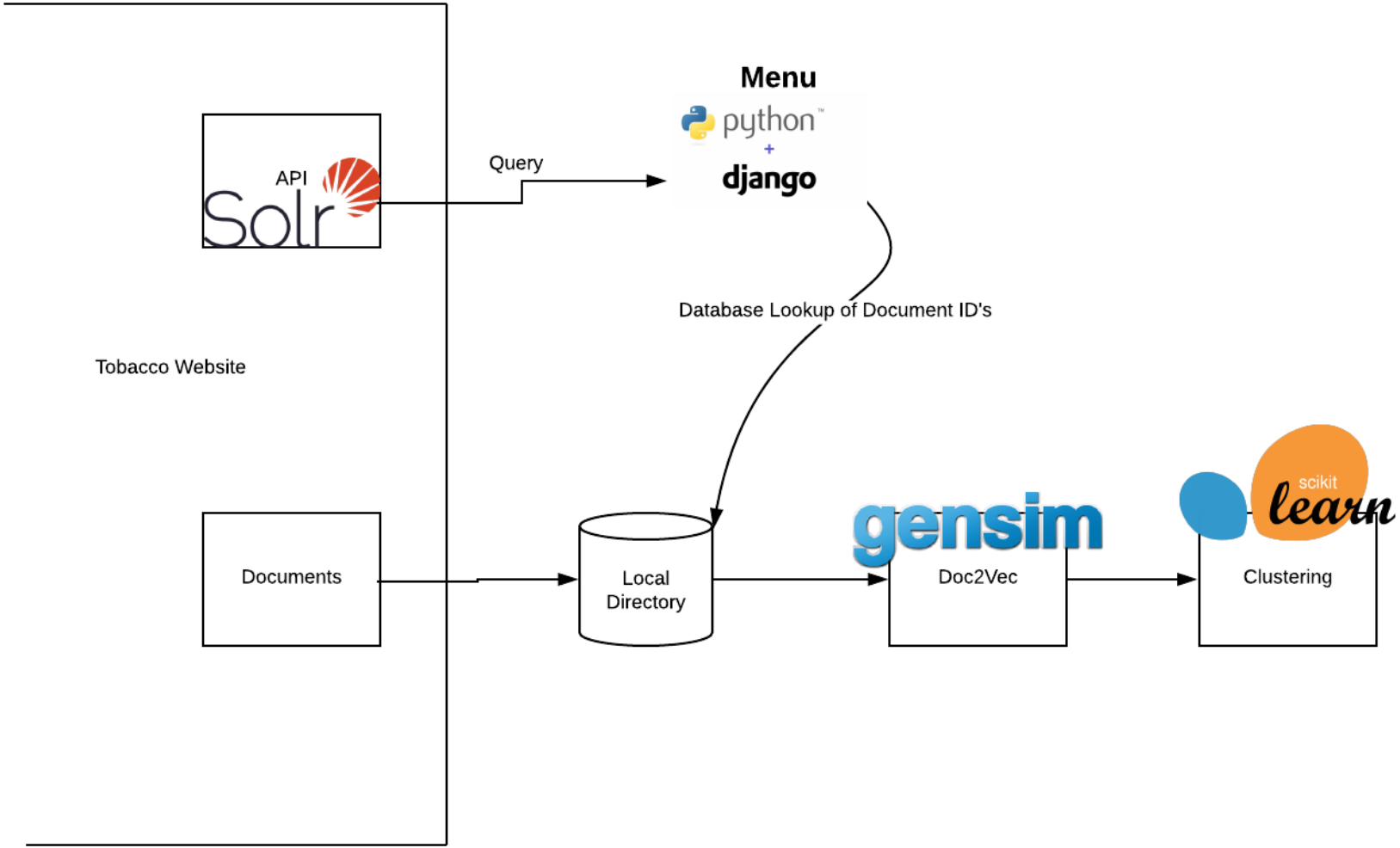
Outline

- Background
- Menu
- Document Gathering
- Doc2Vec
- Similarities
- Clustering
- Lessons Learned
- Acknowledgements
- References
- Questions

Background



- How can we identify ways of combating Big Tobaccos strategies


Menu
python
+
django
API
Solr
Query
Database Lookup of Document ID's
Tobacco Website
Documents
Local
Directory
gensim
Doc2Vec
scikit
learn
Clustering

Menu

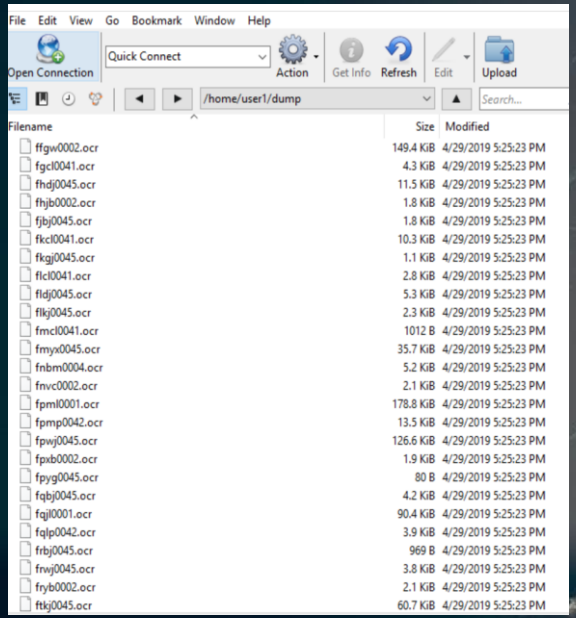
- We stopped work on this portion of the project.
- As it stands, users can select a topic to query the API.
- Working well with storing and finding documents.

Document Gathering

- Downloading Documents
- Organization
- Feeding into the Model

```python
import sys
import shutil
import os

def getDocs(docIdList):

    dirpath = os.getcwd()
    print(dirpath)
    pathLoc = '/home/user1/dump'
    shutil.rmtree(pathLoc)

    try:
        if not os.path.exists(pathLoc):
            os.makedirs(pathLoc)
    except OSError:
        print('Error creating directory')

    j = 0
    for docId in docIdList:
        print(docId)
        id1 = docId[0]
        id2 = docId[1]
        id3 = docId[2]
        id4 = docId[3]

        path = '/home/user1/root/' + id1 + '/' + id2 + '/' + id3 + '/' + id4 + '/' + docId
        print(path)

        exists = os.path.isfile(path + "/" + docId + ".ocr")
        if exists:
            shutil.copyfile(path + "/" + docId + ".ocr", pathLoc + "/" + docId + ".ocr")
        else:
            print(path + " not found")

        #docFile = open(path + "/" + docId + ".ocr", "r")
        #docFile.close()

docList = ['njmp0006', 'txhx0228', 'zznn0003']
getDocs(docList)
```

```
/home/user1
njmp0006
/home/user1/root/n/j/m/p/njmp0006
txhx0228
/home/user1/root/t/x/h/x/txhx0228
/home/user1/root/t/x/h/x/txhx0228 not found
zznn0003
/home/user1/root/z/z/n/n/zznn0003
```



Doc2Vec

```
#returns the inferred paragraph vector for the new document
f = open('/home/user1/corpusTotalVocabTokenized.txt', 'r')
contents = word_tokenize(f.read())
f.close()

i = 0
v1 = model.infer_vector(contents)
#print("V1_infer", v1)



# to find most similar doc using tags
#most_similar finds the top-N most similar words. Positive words contribute positively
#towards the similarity, negative words contribute negatively.
#docvecs hold all trained vectors for the document tags seen during training
#use topn= to specify how many docs you want
similar_doc = model.docvecs.most_similar('42', topn=5)
print(similar_doc)


# to find vector of doc in training data using tags or in other words, printing the vector of document at index 1 in training
#print(model.docvecs['1'])
```

```
[('41', 0.9971927404403687), ('45', 0.994623064994812), ('43', 0.9920123815536499), ('44', 0.9858886003494263), ('39', 0.984
7745299339294)]
```

Doc2Vec Continued

ECKEL & VAUGHAN

MEMORANDUM

TO: RAIS RO Team
FROM: E&V Team
RE: Monthly Analysis: August 27 – September 30, 2015
DATE: October 2, 2015

The following report contextualizes events and media coverage relating to the regulation of various tobacco products during the past month.

Key Takeaways:

- RAI was the subject of action taken by the FDA in two instances this month. On August 27, the FDA issued a warning letter to Santa Fe Natural Tobacco Company for describing Natural American Spirit cigarettes as "natural" and "additive free." On September 15, the FDA issued not substantially equivalent orders that will stop the further sale and distribution of four R.J. Reynolds products: Camel Crush Bold, Pall Mall Deep Set Recessed Filter, Pall Mall Deep Set Recessed Filter Menthol, and Vantage Tech 13. The agency stated that R.J. Reynolds failed to demonstrate that increased yields of harmful or potentially harmful constituents, higher levels of menthol, and/or the addition of new ingredients do not raise different questions of public health. We identified more than 400 articles published regarding the August 27 warning letter, and more than 800 articles published discussing the September 15 NSE orders. (See page 3.)
- There was also a significant amount of guidance and rules issued by the FDA. We saw

ECKEL & VAUGHAN

MEMORANDUM

TO: RAIS RO Team
FROM: E&V Team
RE: Monthly Analysis: August 27 – September 30, 2015
DATE: October 2, 2015

The following report contextualizes events and media coverage relating to the regulation of various tobacco products during the past month.

Key Takeaways:

- RAI was the subject of action taken by the FDA in two instances this month. On August 27, the FDA issued a warning letter to Santa Fe Natural Tobacco Company for describing Natural American Spirit cigarettes as "natural" and "additive free." On September 15, the FDA issued not substantially equivalent orders that will stop the further sale and distribution of four R.J. Reynolds products: Camel Crush Bold, Pall Mall Deep Set Recessed Filter, Pall Mall Deep Set Recessed Filter Menthol, and Vantage Tech 13. The agency stated that R.J. Reynolds failed to demonstrate that increased yields of harmful or potentially harmful constituents, higher levels of menthol, and/or the addition of new ingredients do not raise different questions of public health. We identified more than 400 articles published regarding the August 27 warning letter, and more than 800 articles published discussing the September 15 NSE orders. (See page 3.)
- There was also a significant amount of guidance and rules issued by the FDA. We saw

Similarities

- Using the gensim similarities.index package
- This package uses Annoy (Approximate Nearest Neighbors Oh Yeah)
  - Finds nearest neighbors in vector tree from Doc2Vec model

```
from gensim.similarities.index import AnnoyIndexer

# 100 trees are being used
annoy_index = AnnoyIndexer(model, 100)

# Derive the vector for the words "surgeon general" in our model
test_data = word_tokenize("surgeon general".lower())

vector = model.infer_vector(test_data)
# The instance of AnnoyIndexer we just created is passed
# approximate_neighbors contains a list of 5 docuemnts with the most similarites to the words 'surgeon general'
approximate_neighbors = model.most_similar([vector], topn=5, indexer=annoy_index)

# we now have a list of the 5 nearest neighbors relating to the words surgeon general
```

Clustering

- We can cluster our document vectors using the Scikit-learn Birch clustering algorithm
- Using the Doc2Vec model, we create a list of vectors with each corresponding to a document
- We are returned in order the cluster that each corresponding document is in

```
for d in test_docs:
    X.append(model.infer_vector(d, alpha=start_alpha, steps=infer_epoch))
#


print(X[0])
print(len(X))

from sklearn.cluster import Birch

brc = Birch(branching_factor=50, n_clusters=5, threshold=0.1, compute_labels=True)
brc.fit(X)

clusters = brc.predict(X)
```

Lessons Learned

- Planning is ESSENTIAL and can help the project run much smoother
- Seek advice from the resources that are available
- Working face to face with group members was more productive

Acknowledgements

- Client: David M. Townsend, Ph.D
- Technical Assistance Contacts: Saurabh Chakravarty and Professor Fox

References

1. Glantz, Stanton. *Industry Documents Library*, 2002, www.industrydocumentslibrary.ucsf.edu/tobacco/.
2. “Google Code Archive - Long-Term Storage for Google Code Project Hosting.” *Google*, Google, 29 July 2013, code.google.com/archive/p/word2vec/.
3. “Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.” *ACM Digital Library*, Association for Computing Machinery and Morgan & Claypool, 2016, dl.acm.org/citation.cfm?id=2915031.
4. Gensim: Topic modelling for humans. (n.d.). Retrieved April 25, 2019, from https://radimrehurek.com/gensim/models/doc2vec.html
5. KDnuggets. (n.d.). Retrieved April 25, 2019, from https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-cbow.html
1. Doc2Vec Document Vectorization and clustering. (n.d.). Retrieved April 25, 2019, from http://techscouter.blogspot.com/2018/09/doc2vec-document-vectorization-and.html
2. Text Clustering with doc2vec Word Embedding Machine Learning Model. (2018, October 04). Retrieved April 25, 2019, from http://ai.intelligentonlinetools.com/ml/text-clustering-doc2vec-word-embedding-machine-learning/
3. Qaiser, R. (2017, May 12). How to Extract Words from PDFs with Python. Retrieved April 25, 2019, from https://medium.com/@rqaiserr/how-to-convert-pdfs-into-searchable-key-words-with-python-85aab86c544f
4. Mishra, D. (2018, March 17). DOC2VEC gensim tutorial. Retrieved April 25, 2019, from https://medium.com/mishra.thedeepak/doc2vec-simple-implementation-example-df2afbbfbad5

Used this source in the creation of the Doc2Vec Model

1. Doc2vec tutorial. (n.d.). Retrieved April 25, 2019, from https://rare-technologies.com/doc2vec-tutorial/

Questions?