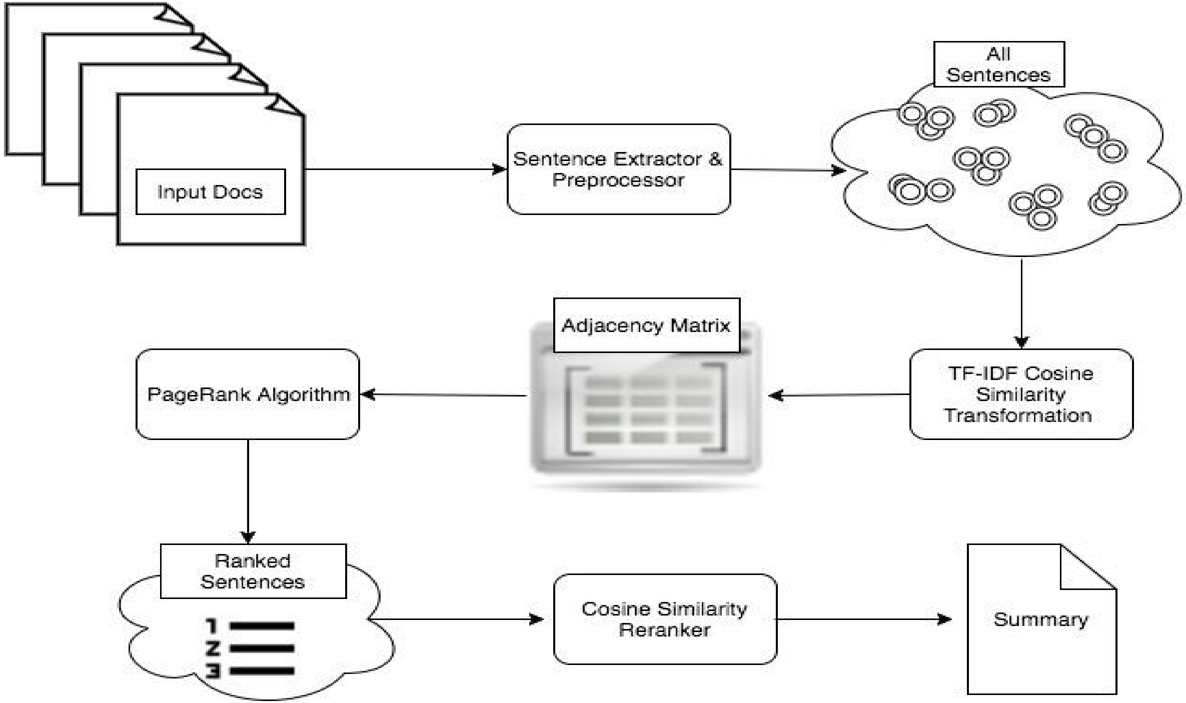
Multi-Document Summarization

Eslam Elsawy, Audrey Holmes,
Masha Ivenskaya

System Architecture


Input Docs
Sentence Extractor & Preprocessor
All Sentences
TF-IDF Cosine Similarity Transformation
Adjacency Matrix
PageRank Algorithm
Ranked Sentences
1
2
3
Cosine Similarity Reranker
Summary

Dataset Pre-Processing

Issues:

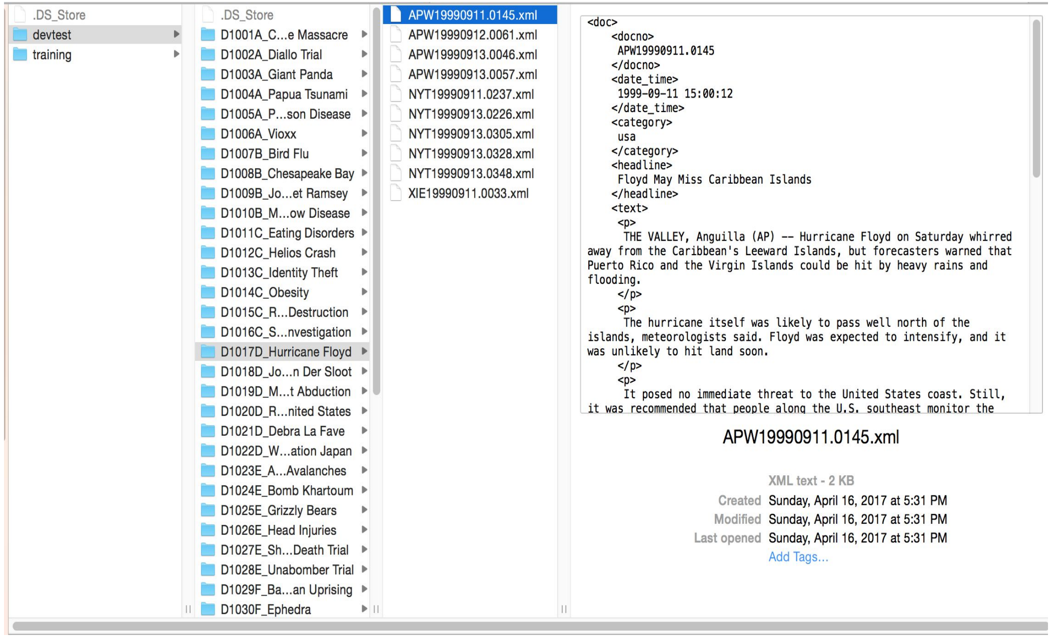
- Non-uniform file naming scheme
- Encoded characters (e.g: &amp)
- Not rooted xml files

What we did:

- Separate cleaning module

Gain:

- Easier to explore the dataset
- Better running times



.DS_Store
devtest
training

.DS_Store
D1001A_C...e Massacre
D1002A_Diallo Trial
D1003A_Giant Panda
D1004A_Papua Tsunami
D1005A_P...son Disease
D1006A_Vioxx
D1007B_Bird Flu
D1008B_Chesapeake Bay
D1009B_Jo...et Ramsey
D1010B_M...ow Disease
D1011C_Eating Disorders
D1012C_Helios Crash
D1013C_Identity Theft
D1014C_Obesity
D1015C_R...Destruction
D1016C_S...nvestigation
D1017D_Hurricane Floyd
D1018D_Jo...n Der Sloot
D1019D_M...t Abduction
D1020D_R...nited States
D1021D_Debra La Fave
D1022D_W...ation Japan
D1023E_A...Avalanches
D1024E_Bomb Khartoum
D1025E_Grizzly Bears
D1026E_Head Injuries
D1027E_Sh...Death Trial
D1028E_Unabomber Trial
D1029F_Ba...an Uprising
D1030F_Ephedra

APW19990911.0145.xml
APW19990912.0061.xml
APW19990913.0046.xml
APW19990913.0057.xml
NYT19990911.0237.xml
NYT19990913.0226.xml
NYT19990913.0305.xml
NYT19990913.0328.xml
NYT19990913.0348.xml
XIE19990911.0033.xml

```
<doc>
    <docno>
     APW19990911.0145
    </docno>
    <date_time>
     1999-09-11 15:00:12
    </date_time>
    <category>
     usa
    </category>
    <headline>
     Floyd May Miss Caribbean Islands
    </headline>
    <text>
     <p>
      THE VALLEY, Anguilla (AP) -- Hurricane Floyd on Saturday whirred
away from the Caribbean's Leeward Islands, but forecasters warned that
Puerto Rico and the Virgin Islands could be hit by heavy rains and
flooding.
     </p>
     <p>
      The hurricane itself was likely to pass well north of the
islands, meteorologists said. Floyd was expected to intensify, and it
was unlikely to hit land soon.
     </p>
     <p>
      It posed no immediate threat to the United States coast. Still,
it was recommended that people along the U.S. southeast monitor the
```

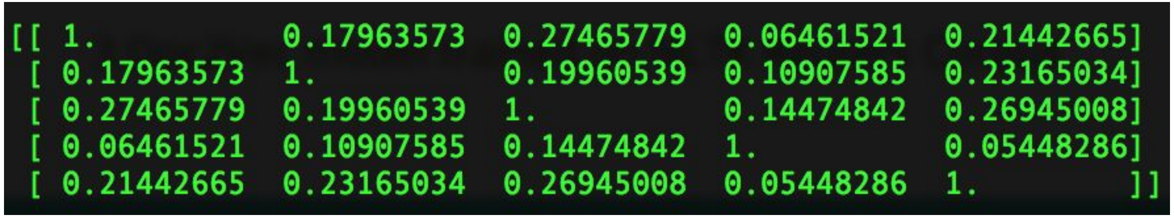
APW19990911.0145.xml

XML text - 2 KB

Created Sunday, April 16, 2017 at 5:31 PM
Modified Sunday, April 16, 2017 at 5:31 PM
Last opened Sunday, April 16, 2017 at 5:31 PM

Content Selection - LexRank - Sentence Similarity

- All sentences → adjacency matrix by similarity score



```
[[ 1.          0.17963573  0.27465779  0.06461521  0.21442665]
 [ 0.17963573  1.          0.19960539  0.10907585  0.23165034]
 [ 0.27465779  0.19960539  1.          0.14474842  0.26945008]
 [ 0.06461521  0.10907585  0.14474842  1.          0.05448286]
 [ 0.21442665  0.23165034  0.26945008  0.05448286  1.        ]]
```

- Approaches to measure pairwise similarity:

  1.Cosine Similarity

  2.Cosine Similarity with TF-IDF (highest Rouge scores)

  3.Doc2Vec model trained on NLTK Reuters Corpus

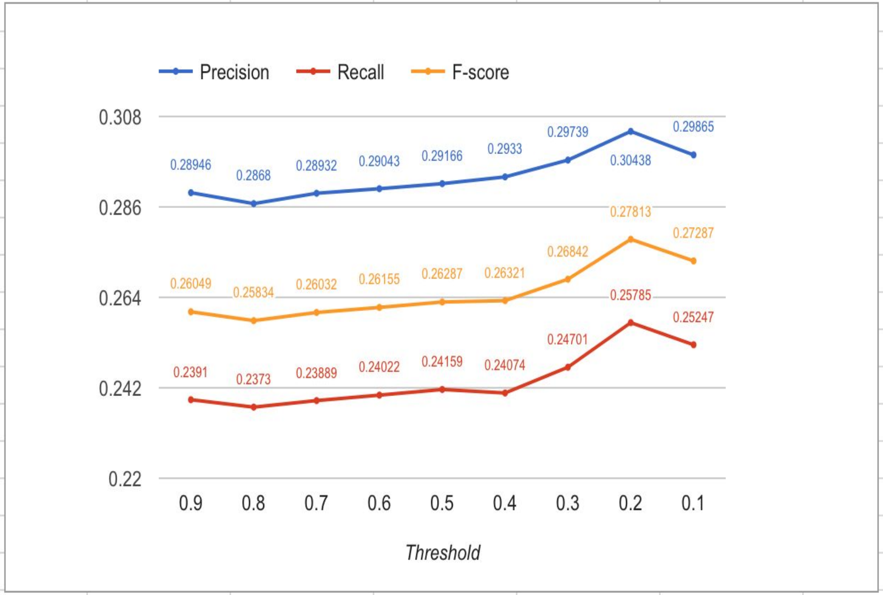
Content Selection - LexRank - Algorithm

- Binarize similarity matrix ***M*** with 0.15 threshold.
- Implement LexRank using Power Method:

1. Initialize $\boldsymbol{p_0}$ vector with uniform distribution.
2. Iteratively update $\boldsymbol{p_t}$ such that $\boldsymbol{p_t} = \boldsymbol{M^T p_{t-1}}$ until $||\boldsymbol{p_t} - \boldsymbol{p_{t-1}}||$ is sufficiently small.
3. Return $\boldsymbol{p_t}$.

Information Ordering & Content Realization

- Input: sentences sorted by score
- TF-IDF cosine similarity ordering [1]
- One Sentence per line
- Max 100 words



Threshold?

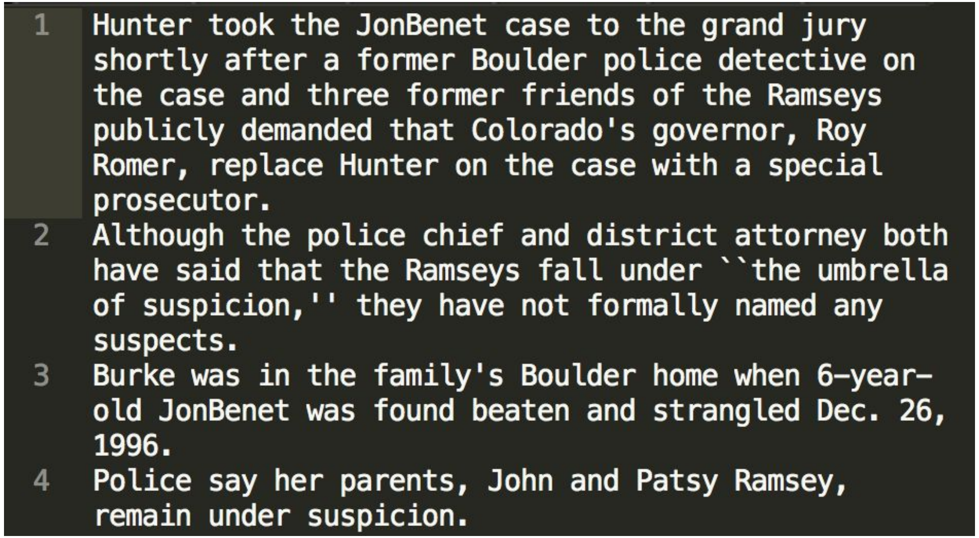
- At 1.0 => R-1 R = 0.239
- At 0.2 => R-1 R = 0.258 (+0.019)

Results

- Runtime ~= 5 - 10 seconds
- Best results:
  - Content Selection: TF-IDF cos sim
  - Ordering: threshold = 0.2

ROUGE-L	Recall
ROUGE-1	0.25785
ROUGE-2	0.07108
ROUGE-3	0.02438
ROUGE-4	0.00847

Sample output
Topic: JonBenet Ramsey Murder



```
Hunter took the JonBenet case to the grand jury shortly after a former Boulder police detective on the case and three former friends of the Ramseys publicly demanded that Colorado's governor, Roy Romer, replace Hunter on the case with a special prosecutor.
Although the police chief and district attorney both have said that the Ramseys fall under ``the umbrella of suspicion,'' they have not formally named any suspects.
Burke was in the family's Boulder home when 6-year-old JonBenet was found beaten and strangled Dec. 26, 1996.
Police say her parents, John and Patsy Ramsey, remain under suspicion.
```

Issues and Successes

Success: Implementing tf-idf raised ROUGE-1 recall from 0.13 to 0.26.

Problems that need fixing:

- Meta-info like:
    - LITTLETON, Colo. (AP) --
    - NEW YORK _ The parents of Amadou Diallo plan to meet with the Bronx district
- Quotes without attributions
- Pronouns without reference
- Ordering of the sentences within summary
- Long sentences

References

[1] Radev, Dragomir R., et al. "MEAD-A Platform for Multidocument Multilingual Text Summarization." *LREC*. 2004.

[2] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research 22 (2004): 457-479.

[3] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. 2004.

Questions ?