News Text Summarization

•••

Jerome Tan

Business Overview

Content being published is increasing everyday



• To make use of automatic text summarization on the latest news articles to create short and concise summaries in order to save time on reading.

Technical Overview

- Web Scraping
- Importing and Data Cleaning
- Exploratory Data Analysis
- Modelling
- Evaluation
- Conclusion

Data Set

- Web Scraping News Sources:
 - Associated Press News
 - o BBC
 - o NPR

Kaggle Data Set:

Contains articles & summaries from inshorts

o Feb - Aug 2017

Trump's defense shifts to not 'impeachable' even if true

By LISA MASCARO, ERIC TUCKER and ZEKE MILLER toda



China counts 170 virus deaths, new countries find infections



By KEN MORITSUG

10 Things to Know for Today



By The Associated Press an hour ago

Biden under pressure to prove he can thwart new GOP attacks



By BILL BARROV today

Do masks offer protection from new virus? It depends



By CARLA K. JOHNSO today Wary of irking China, Trump offers rosy take on virus threat



By DARLENE SUPERVILLE, AAMER MADHANI and JONATHAN LEMIRE

Vanessa Bryant statement: 'We are completely devastated'



By TIM REYNOLDS today

Brain injuries in Iraq put attention on invisible war wounds

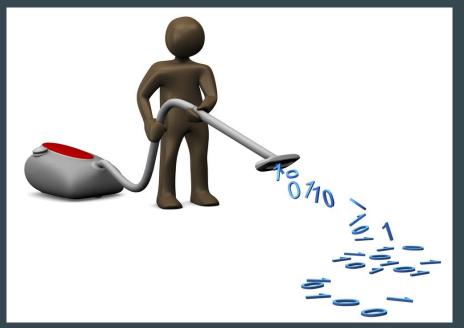


By ROBERT BURNS

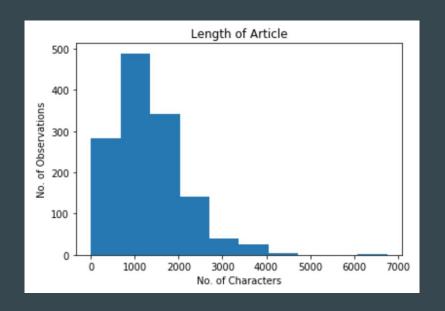
Brexit finally arrives Friday: A momentous yet quiet moment

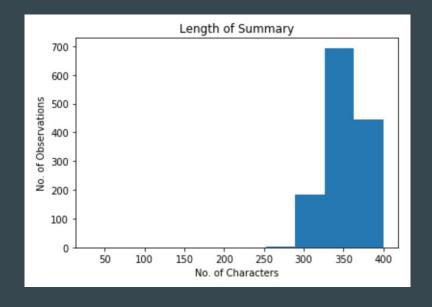
Data Cleaning

- Mainly on Kaggle Data Set
 - Non-alphanumic characters
 - Hashtags
 - Url links

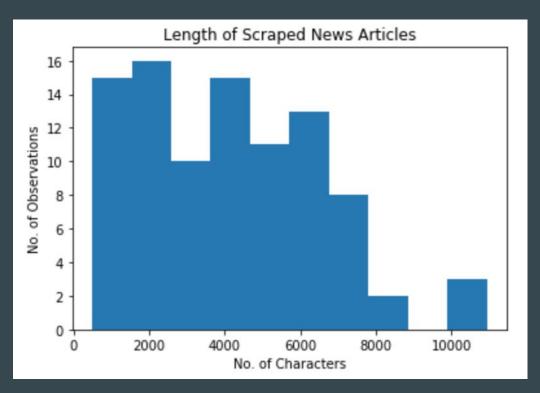


Exploratory Data Analysis





Exploratory Data Analysis



Modelling

- Lede-3 (Baseline Model)
- LexRank
- Luhn
- LSA
- TextRank
- Edmundson
- GloVe + TextRank

Evaluation

- Evaluation is based on Rouge Score
 - ROUGE-n recall=40% means that 40% of the n-grams in the reference summary are also present in the generated summary.
 - PROUGE-n precision=40% means that 40% of the n-grams in the generated summary are also present in the reference summary
 - ROUGE-n F1-score=40% is more difficult to interpret, like any F1-score.

	lede_3	lexrank	luhn	Isa	textrank	edmundson	glove_textrank
recall	0.513883	0.465951	0.494760	0.376341	0.483477	0.376341	0.472039
precision	0.397336	0.327083	0.282885	0.298557	0.270421	0.298557	0.296741
f1	0.439107	0.377744	0.355731	0.327353	0.342800	0.327353	0.359540

Conclusion

- Lede-3 algorithm best performing
- Likely due to nature of how news articles are written
- Other applications for automatic text summarization
 - o Reviews
 - Non-news articles
 - Blogpost

Future Work

- Look into other strategies for text summarization
 - Abstractive text summarization
 - Mixed strategy text summarization
- Multiclass classification for news category
- Named Entity Recognition Sentiment Analysis

Thank You