Project Fletcher: An NLP Analysis of the Oil Crash

By Sunne Kuo

Bloomberg API

-High quality content

-Focused on finance, markets and economics

-Open sourced, free for all to use

Bloomberg API

-High quality content

-Focused on finance, markets and economics

-Open sourced, free for all to use

-Extremely difficult to access the API

OOMBERG AP

-High quality conten

-Focused on markets an
economics

-Open sourced, se

-Extremely difficult to access the API

The New York Times Developers

-High quality content

-Has several sections dedicated to finance, markets, business and economics

-Best of all, easy to access

-However API did not provide full articles, many URL’s were empty

Newspaper - Article Scraping

Pros

-Very fast and easy to use

-Works on 10+ languages

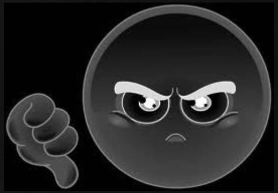
-Full article, keyword and image extraction

-Everything is in Unicode

-NLP package included!

Cons

-hard to install on Python2

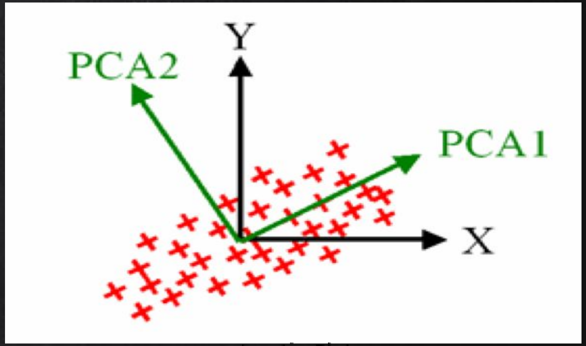


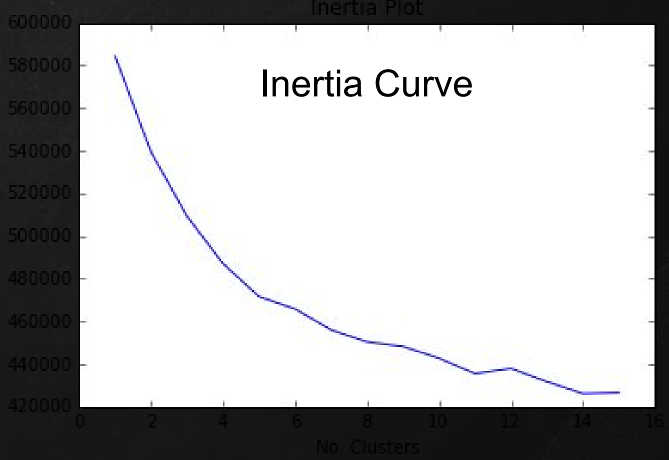
NLP + PCA

-Converted articles to sparse matrix using CountVectorizer

-Ran PCA, cut down the matrix

-Analyzed inertia curve




Inertia Curve

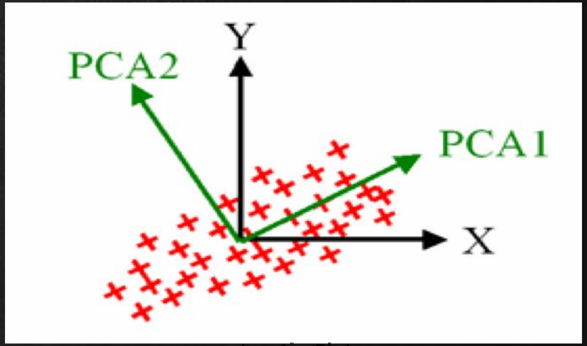
NLP + PCA

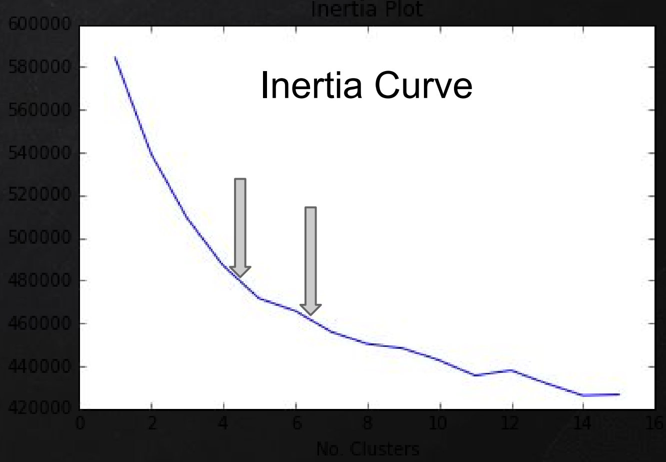
-Converted articles to sparse matrix using CountVectorizer

-Ran PCA, cut down the matrix

-Analyzed inertia curve

-Clustered in 4-6 range





4 – 6

NLP + PCA

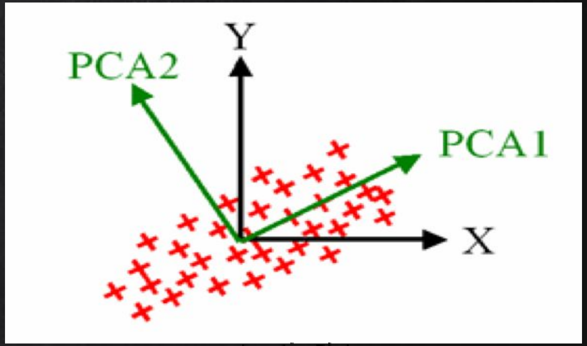
-Converted articles to sparse matrix using CountVectorizer

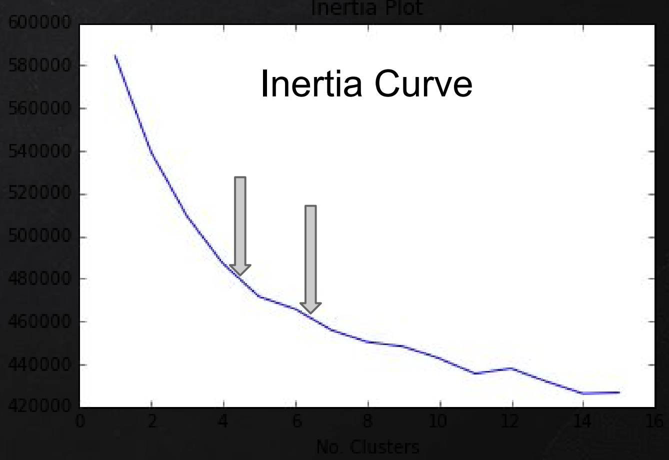
-Ran PCA, cut down the matrix

-Analyzed inertia curve

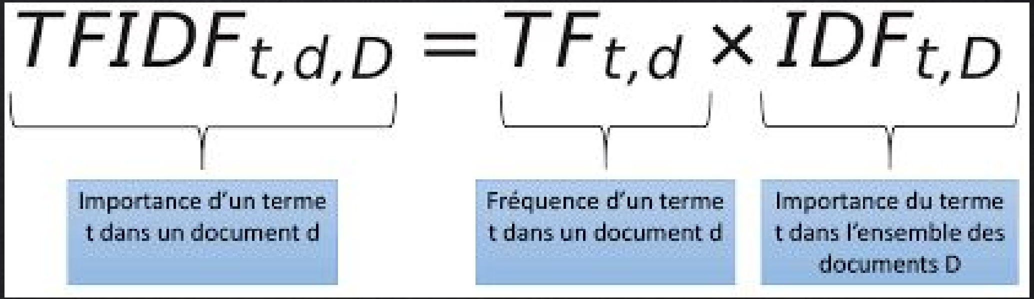
-Clustered in 4-6 range

-Results of Clusters were **ONLY OKAY**





4 – 6



$$\underbrace{TFIDF_{t,d,D}}_{\text{Importance d'un terme t dans un document d}} = \underbrace{TF_{t,d}}_{\text{Fréquence d'un terme t dans un document d}} \times \underbrace{IDF_{t,D}}_{\text{Importance du terme t dans l'ensemble des documents D}}$$

-Words were used frequently over multiple articles

-Tuned TFIDF arguments for better clusters

-Reran k-means clusters and extracted keywords/topics

Next Steps

–Get access to Bloomberg API and scrape those articles

–Dig deeper into topics extracted

–Look exclusively at Op-ed articles for sentiment analysis

–Analyze articles on a time series basis

Thank you