SPARK+AI
SUMMIT EUROPE

Time-Series Anomaly Detection in Plaintext Using Apache Spark

Jerry Schirmer, Ph.D.
SparkCognition

#SAISDS10

SparkCognition's Project Minerva

René-Antoine Houasse

Story of Minerva

Minerva Giving Her Shield to Perseus,

1697

(Public Domain)

Project Minerva

The underlying use cases for Minerva are to take unstructured text, aggregate it, and perform three functions:

Detect anomalous text

- Extract features from unstructured text fields
- Sort feature sets into "normal" and "abnormal"
- Produce original text from day indicative of "normality" of original text

Produce predictive analytics

- Use extracted features to predict time-series data
- Potentially join features with other relevant data

Prioritize text for analysis

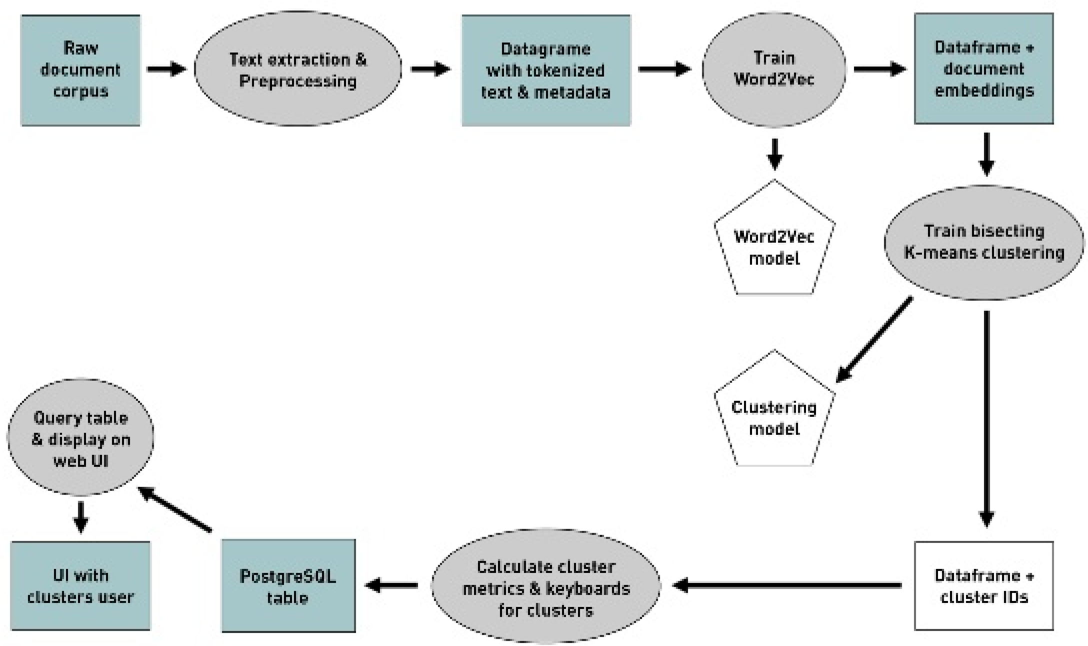
- Leverage models to create reading lists for human users
- Use more traditional ML techniques to tease out statistical relationships on dependent variables
- Use expert knowledge to detect relevant wheat from textual chaff

Where Spark comes in

Why Spark?

- Need to be able to process text at scale
- Need to integrate ML algorithms
- Long-term plans require support for streaming

Anomaly workflow



- Feature extraction, as always is the ML magic
- Word2Vec from Spark worked better than expected
- Algebraic nature of W2V means natural clustering
- Application to time-series

Results

The clusters have natural anomaly detection behavior

```
>>> daily_clusters.groupBy("prediction").count().show()
+----------+-----+
|prediction|count|
+----------+-----+
|        65|    2|
|        61|   29|
|        59|   39|
|        58|    2|
|        68|   92|
|        67|  994|
+----------+-----+
```

More results

- Clustering on named entities created meaningful results
- For instance, when run against news data, we had a cluster with the following entities:
  - *Fed Chair*
  - *Janet Yellin*
  - *Jerome Powell*
  - *Lael Brainard (member of Fed board)*
  - *Edward Nowotny (Governor of Austria's central bank and European Central Bank)*
  - *Haruhiko Kuroda (Governor of the Bank of Japan)*

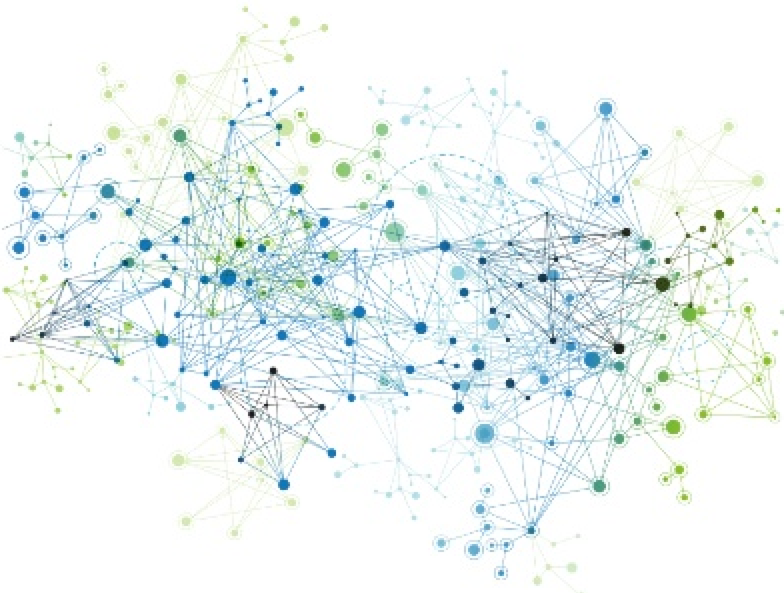
Results — Regression

Performed well with (notoriously difficult) financial data

Built an oil price regression model to predict "high variance" oil days with 55% AUROC

Workflow has potential for user error

Extensions



- Summer interns created module for cluster explainability using topic modeling
- Bisecting k-means is justified, but better metrics for clustering algorithms would make sense
- Different vectorization techniques (LSTM autoencoder, in particular)