Text Mining for Economics and Finance
Lecture 5.2: Word Embeddings Estimation

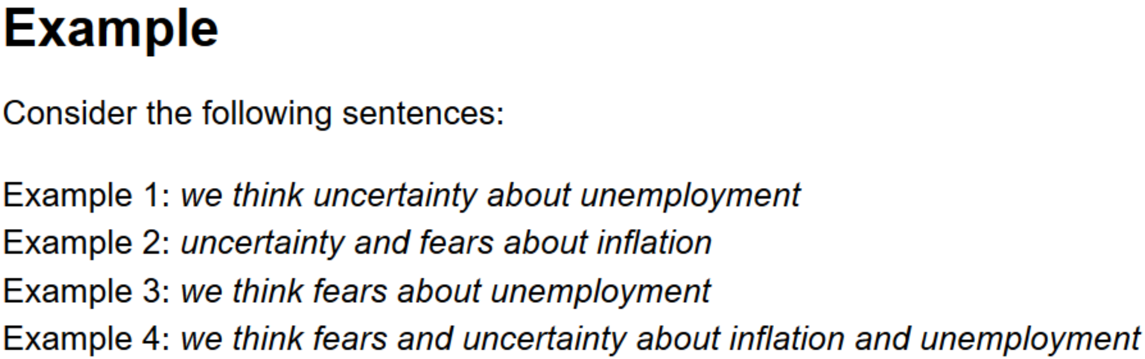
Paul E. Soto, Ph.D.[1]
Robert H. Smith School of Business
University of Maryland

[1] Opinions expressed in this presentation are those of the instructor and not necessarily those of the FDIC

Agenda

- Difference between Skip Gram and CBOW models
- Estimating Word2Vec with Gensim
    - Understand the main parameters such as size, window, sg, min_count
- Reducing the dimensions further of the word embeddings using t-SNE
- Creating relationships or analogies using the **most_similar** method

Recap of Word Embeddings



Example

Consider the following sentences:

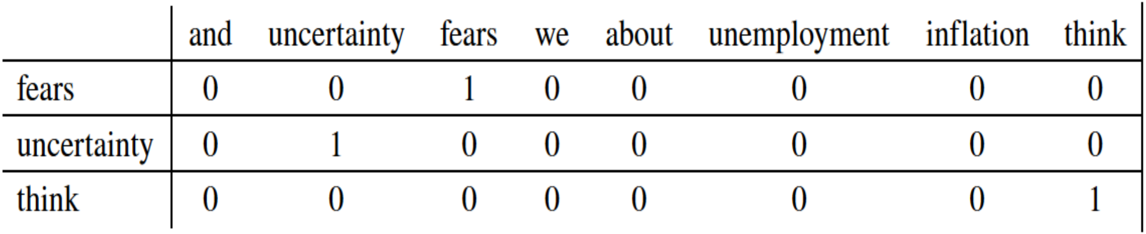
Example 1: *we think uncertainty about unemployment*

Example 2: *uncertainty and fears about inflation*

Example 3: *we think fears about unemployment*

Example 4: *we think fears and uncertainty about inflation and unemployment*

Recap of Word Embeddings



	and	uncertainty	fears	we	about	unemployment	inflation	think
fears	0	0	1	0	0	0	0	0
uncertainty	0	1	0	0	0	0	0	0
think	0	0	0	0	0	0	0	1

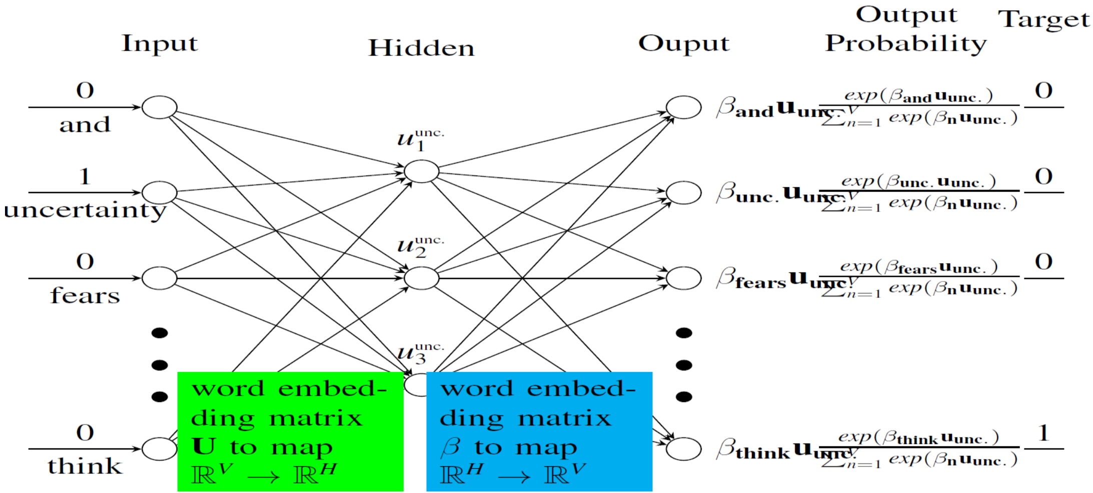
We can represent words with one-hot encoded vectors but this is not useful!

Recap of Word Embeddings

So, we estimate word embeddings

- Unsupervised model
- Creates vector representations of words
- Distances in the vector space represent syntactic and semantic similarities
- Estimated by setting up prediction exercises
    - Skip Gram Model: use target word to predict context words
    - Continuous Bag of Words (CBOW): use context words to predict target
- Let's look at one training exercise....

Recap of Word Embeddings


Input
Hidden
Ouput
Output Probability
Target
0
and
1
uncertainty
0
fears
0
think
$u_1^{\text{unc.}}$
$u_2^{\text{unc.}}$
$u_3^{\text{unc.}}$
$\beta_{\mathbf{and}}\mathbf{u}_{\mathbf{unc.}}$
$\frac{exp(\beta_{\mathbf{and}}\mathbf{u}_{\mathbf{unc.}})}{\sum_{n=1}^{V} exp(\beta_{\mathbf{n}}\mathbf{u}_{\mathbf{unc.}})}$
0
$\beta_{\mathbf{unc.}}\mathbf{u}_{\mathbf{unc.}}$
$\frac{exp(\beta_{\mathbf{unc.}}\mathbf{u}_{\mathbf{unc.}})}{\sum_{n=1}^{V} exp(\beta_{\mathbf{n}}\mathbf{u}_{\mathbf{unc.}})}$
0
$\beta_{\mathbf{fears}}\mathbf{u}_{\mathbf{unc.}}$
$\frac{exp(\beta_{\mathbf{fears}}\mathbf{u}_{\mathbf{unc.}})}{\sum_{n=1}^{V} exp(\beta_{\mathbf{n}}\mathbf{u}_{\mathbf{unc.}})}$
0
$\beta_{\mathbf{think}}\mathbf{u}_{\mathbf{unc.}}$
$\frac{exp(\beta_{\mathbf{think}}\mathbf{u}_{\mathbf{unc.}})}{\sum_{n=1}^{V} exp(\beta_{\mathbf{n}}\mathbf{u}_{\mathbf{unc.}})}$
1

word embedding matrix $\mathbf{U}$ to map $\mathbb{R}^V \longrightarrow \mathbb{R}^H$

word embedding matrix β to map $\mathbb{R}^H \longrightarrow \mathbb{R}^V$

Recap of Word Embeddings

Estimation strategy

- First, create a random **U** and **beta** matrices
- For each word in every sentence, predict the target word (depending on Skip Gram or CBOW)
- Using backpropogation, shift **U** and **beta** matrices to improve prediction
- Do this enough times until predictions meet a threshold

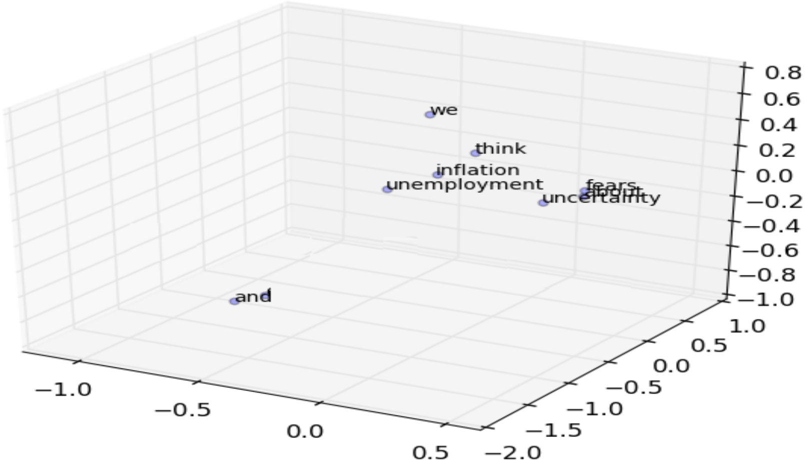
Recap of Word Embeddings

Both $\mathbf{U}$ and β represent the word embeddings

$$\mathbf{U} = \begin{bmatrix} u_1^{and} & u_1^{uncertainty} & u_1^{fears} \cdots & u_1^{think} \\ u_2^{and} & u_2^{uncertainty} & u_2^{fears} \cdots & u_2^{think} \\ u_3^{and} & u_3^{uncertainty} & u_3^{fears} \cdots & u_3^{think} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{and} & \mathbf{u}_{and} & \cdots & \mathbf{u}_{think} \end{bmatrix}$$

Let’s plot these word embeddings...

Recap of Word Embeddings



Application of Word Embeddings

Application of Word Embeddings

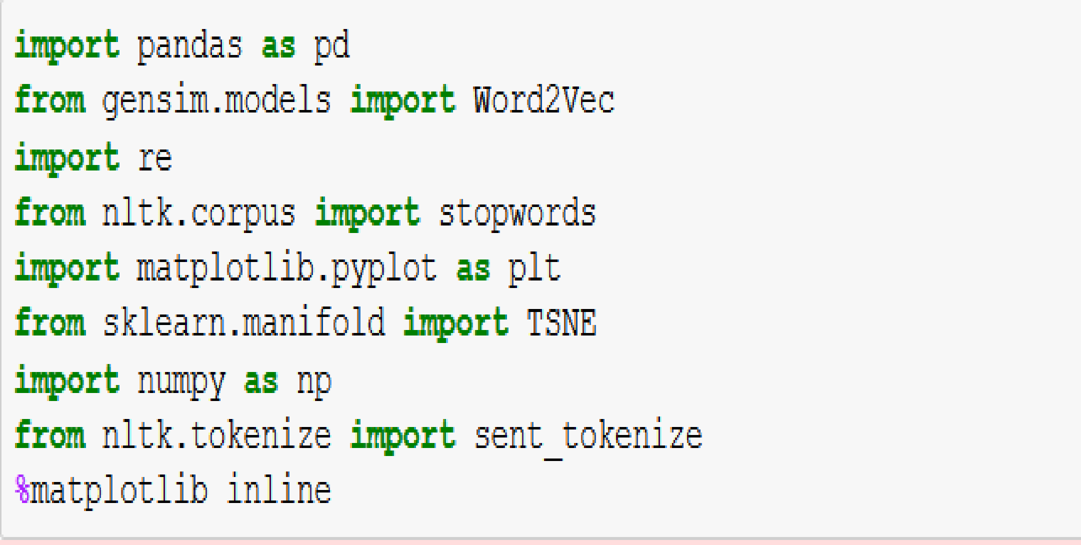
Enron Corporation Accounting Scandal

- American energy company based in Houston, Texas
- Founded in 1985
- Over years, expanded to trading
- Hid financial troubles using dubious accounting loopholes/practices
  - E.g. “mark-to-market accounting"
- December 2, 2001, Enron filed for bankruptcy (nearly $60 billion in assets)
- Scandal led to Sarbanes-Oxley Act in 2002

Dataset

- 500,000 emails generated by Enron employees
- Collected by the Federal Energy Regulatory Commission during investigation
- 0.5% sample randomly drawn for this exercise
- Complete dataset available at https://www.cs.cmu.edu/ ./enron/

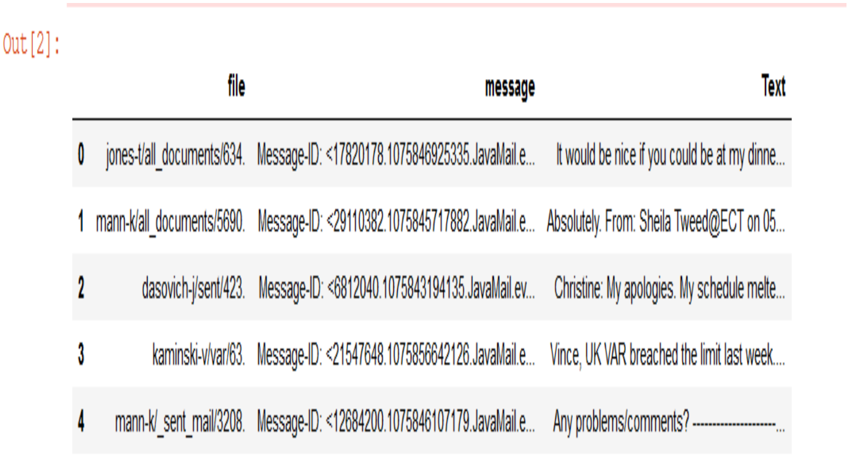
Word Embeddings in Python



```
import pandas as pd
from gensim.models import Word2Vec
import re
from nltk.corpus import stopwords
import matplotlib.pyplot as plt
from sklearn.manifold import TSNE
import numpy as np
from nltk.tokenize import sent_tokenize
%matplotlib inline
```

Word Embeddings in Python

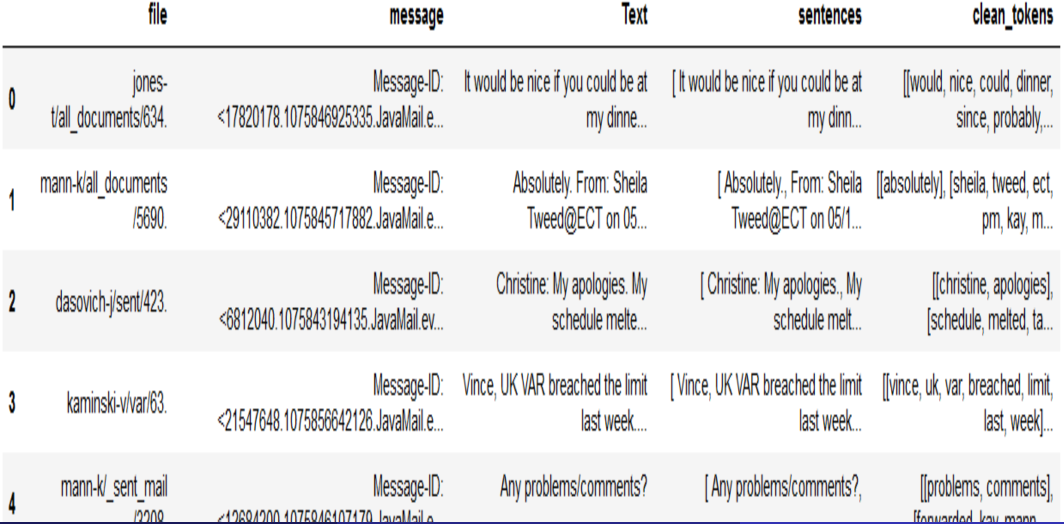
Enron Email Dataset (As published at https://www.cs.cmu.edu/ ./enron/)



Out[2]:

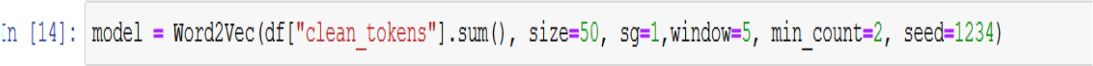
	file	message	Text
0	jones-t/all_documents/634.	Message-ID: <17820178.1075846925335.JavaMail.e...	It would be nice if you could be at my dinne...
1	mann-k/all_documents/5690.	Message-ID: <29110382.1075845717882.JavaMail.e...	Absolutely. From: Sheila Tweed@ECT on 05...
2	dasovich-j/sent/423.	Message-ID: <6812040.1075843194135.JavaMail.ev...	Christine: My apologies. My schedule melte...
3	kaminski-v/var/63.	Message-ID: <21547648.1075856642126.JavaMail.e...	Vince, UK VAR breached the limit last week....
4	mann-k/_sent_mail/3208.	Message-ID: <12684200.1075846107179.JavaMail.e...	Any problems/comments? --------------------...

Word Embeddings in Python



	file	message	Text	sentences	clean_tokens
0	jones-t/all_documents/634.	Message-ID: <17820178.1075846925335.JavaMail.e...	It would be nice if you could be at my dinne...	[It would be nice if you could be at my dinn...	[[would, nice, could, dinner, since, probably,...
1	mann-k/all_documents/5690.	Message-ID: <29110382.1075845717882.JavaMail.e...	Absolutely. From: Sheila Tweed@ECT on 05...	[Absolutely., From: Sheila Tweed@ECT on 05/1...	[[absolutely], [sheila, tweed, ect, pm, kay, m...
2	dasovich-j/sent/423.	Message-ID: <6812040.1075843194135.JavaMail.ev...	Christine: My apologies. My schedule melte...	[Christine: My apologies., My schedule melt...	[[christine, apologies], [schedule, melted, ta...
3	kaminski-v/var/63.	Message-ID: <21547648.1075856642126.JavaMail.e...	Vince, UK VAR breached the limit last week....	[Vince, UK VAR breached the limit last week...	[[vince, uk, var, breached, limit, last, week]...
4	mann-k/_sent_mail /3208.	Message-ID: <12684200.1075846107179.JavaMail.e	Any problems/comments?	[Any problems/comments?,	[[problems, comments], [forwarded, kay, mann

Word Embeddings in Python

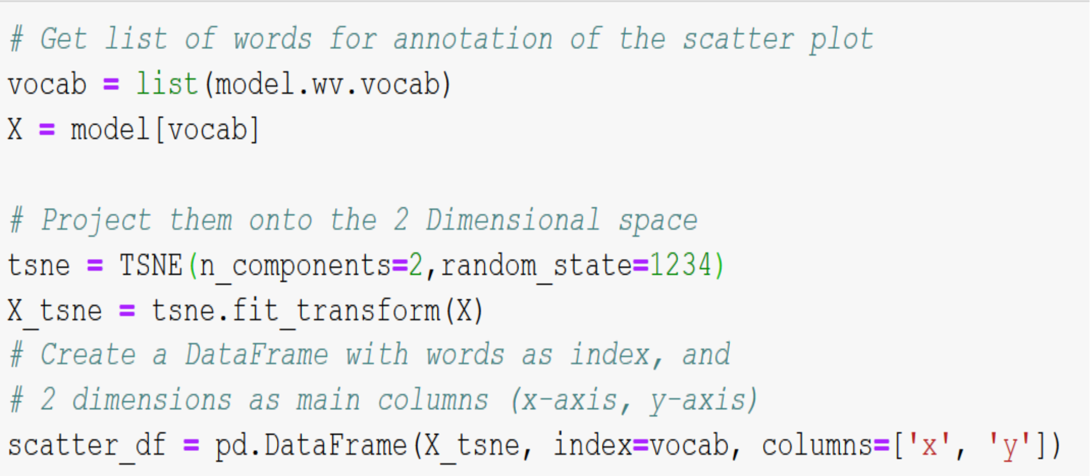


```
In [14]: model = Word2Vec(df["clean_tokens"].sum(), size=50, sg=1,window=5, min_count=2, seed=1234)
```

- size: dimension of the word embeddings
- sg: whether or not you want to estimate the Skip Gram model or the CBOW
- min_count: Ignores all words with total frequency lower than this
- window: The maximum distance to use for predictions

Word Embeddings in Python

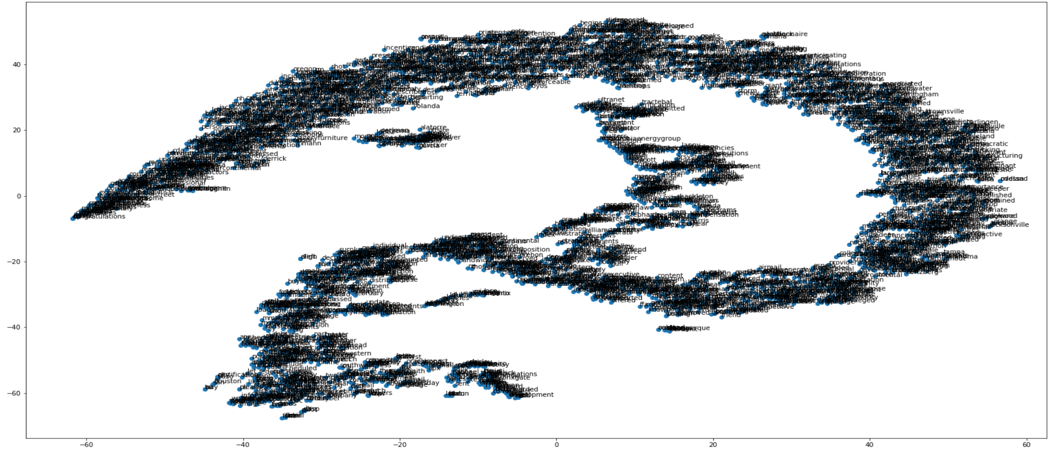
Dimension Reduction using t-SNE



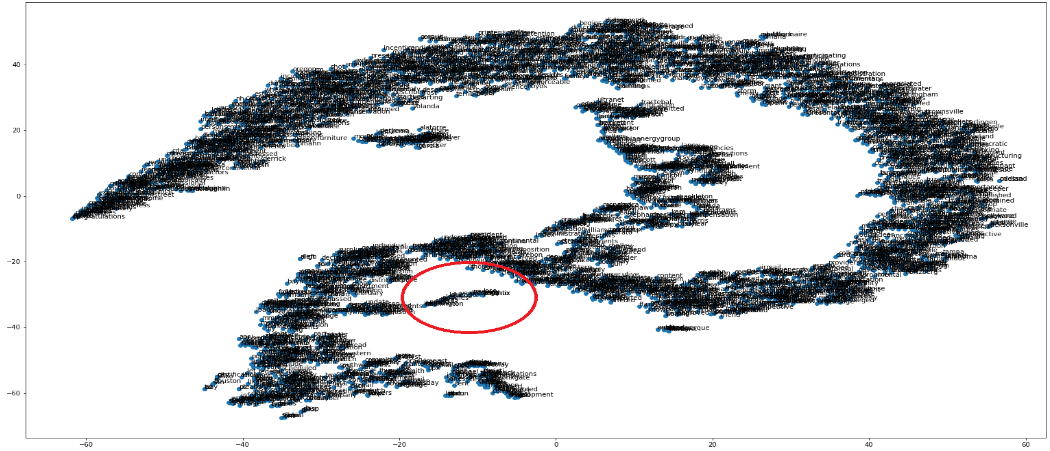
```
# Get list of words for annotation of the scatter plot
vocab = list(model.wv.vocab)
X = model[vocab]

# Project them onto the 2 Dimensional space
tsne = TSNE(n_components=2,random_state=1234)
X_tsne = tsne.fit_transform(X)
# Create a DataFrame with words as index, and
# 2 dimensions as main columns (x-axis, y-axis)
scatter_df = pd.DataFrame(X_tsne, index=vocab, columns=['x', 'y'])
```

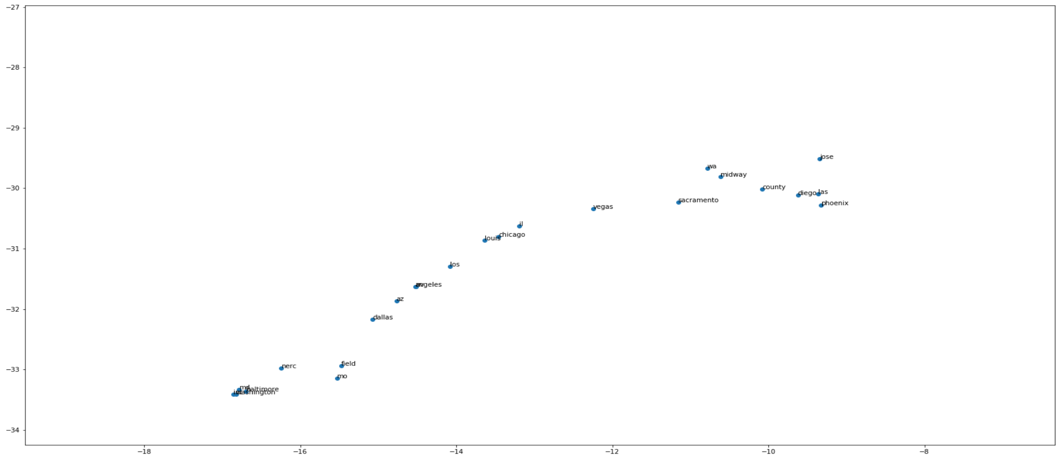
Word Embeddings in Python



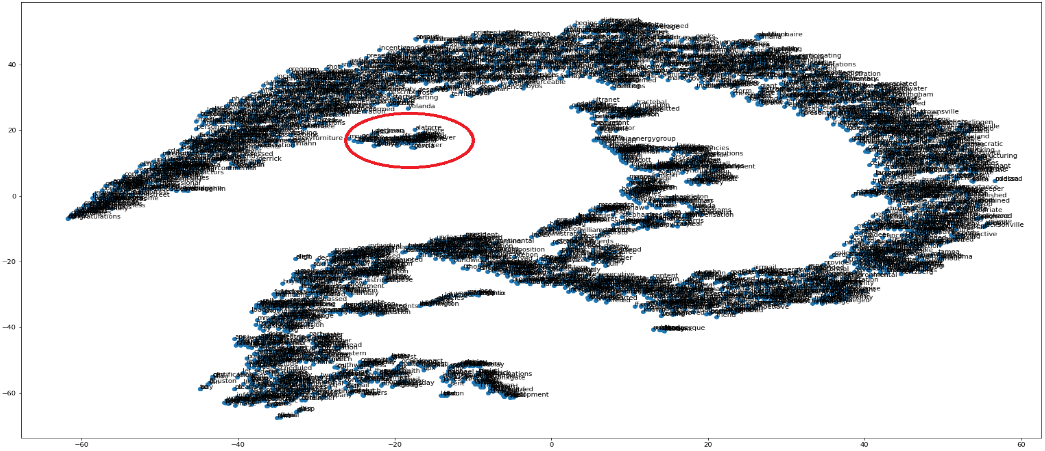
Word Embeddings in Python



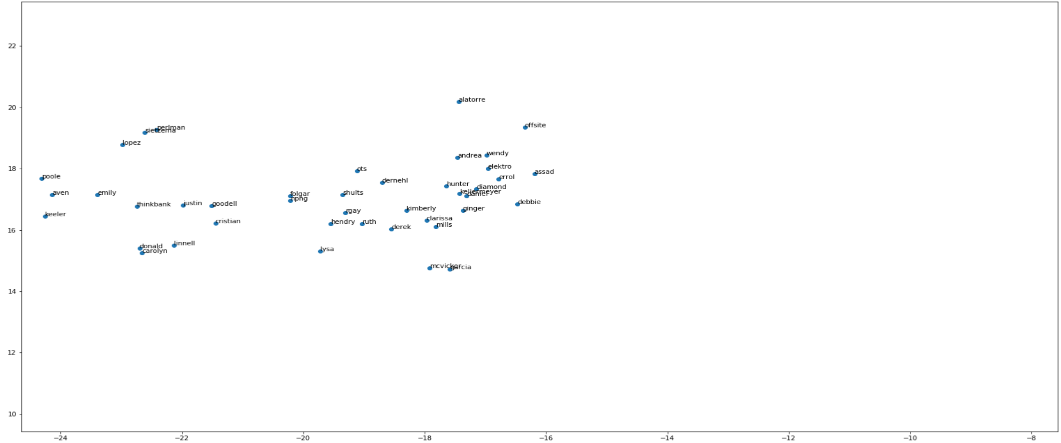
Word Embeddings in Python



Word Embeddings in Python



Word Embeddings in Python



Word Embeddings in Python

Relationships and Analogies in H-Dimensional space

- $vec(vegas) - vec(nv) + vec(phoenix) = ?$

Word Embeddings in Python

Relationships and Analogies in H-Dimensional space

- $vec(vegas) - vec(nv) + vec(phoenix) = ?$
  - $vec(az)$
- $vec(az) - vec(phoenix) + vec(mo) = ?$

Word Embeddings in Python

Relationships and Analogies in H-Dimensional space

- $vec(vegas) - vec(nv) + vec(phoenix) = ?$
    - $vec(az)$
- $vec(az) - vec(phoenix) + vec(mo) = ?$
    - $vec(louis)$

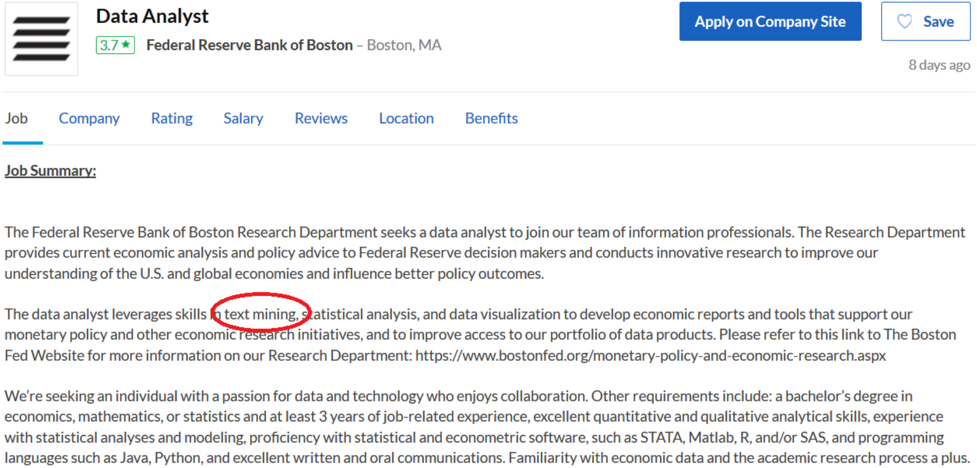
Now What?

Am I ever going to use this in "real" life?

Text Mining is Part of the Future

Today’s Gold = Data

Text Mining Jobs



Data Analyst

3.7 ★ **Federal Reserve Bank of Boston** – Boston, MA

8 days ago

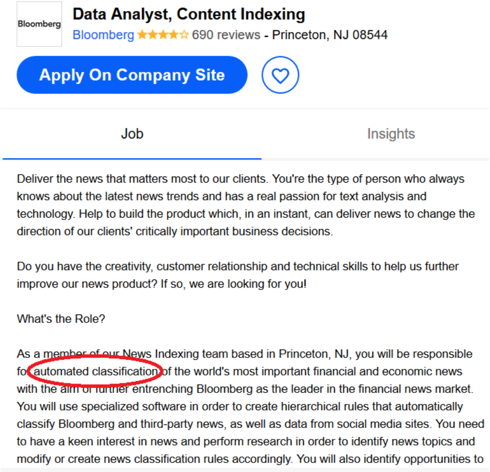
Job Summary:

The Federal Reserve Bank of Boston Research Department seeks a data analyst to join our team of information professionals. The Research Department provides current economic analysis and policy advice to Federal Reserve decision makers and conducts innovative research to improve our understanding of the U.S. and global economies and influence better policy outcomes.

The data analyst leverages skills in text mining, statistical analysis, and data visualization to develop economic reports and tools that support our monetary policy and other economic research initiatives, and to improve access to our portfolio of data products. Please refer to this link to The Boston Fed Website for more information on our Research Department: https://www.bostonfed.org/monetary-policy-and-economic-research.aspx

We're seeking an individual with a passion for data and technology who enjoys collaboration. Other requirements include: a bachelor's degree in economics, mathematics, or statistics and at least 3 years of job-related experience, excellent quantitative and qualitative analytical skills, experience with statistical analyses and modeling, proficiency with statistical and econometric software, such as STATA, Matlab, R, and/or SAS, and programming languages such as Java, Python, and excellent written and oral communications. Familiarity with economic data and the academic research process a plus.

Text Mining Jobs



Data Analyst, Content Indexing

Bloomberg ★★★★☆ 690 reviews - Princeton, NJ 08544

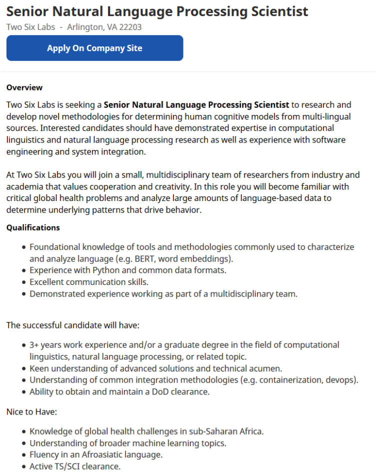
Deliver the news that matters most to our clients. You're the type of person who always knows about the latest news trends and has a real passion for text analysis and technology. Help to build the product which, in an instant, can deliver news to change the direction of our clients' critically important business decisions.

Do you have the creativity, customer relationship and technical skills to help us further improve our news product? If so, we are looking for you!

What's the Role?

As a member of our News Indexing team based in Princeton, NJ, you will be responsible for automated classification of the world's most important financial and economic news with the aim of further entrenching Bloomberg as the leader in the financial news market. You will use specialized software in order to create hierarchical rules that automatically classify Bloomberg and third-party news, as well as data from social media sites. You need to have a keen interest in news and perform research in order to identify news topics and modify or create news classification rules accordingly. You will also identify opportunities to

Text Mining Jobs



Senior Natural Language Processing Scientist

Two Six Labs - Arlington, VA 22203

Overview

Two Six Labs is seeking a **Senior Natural Language Processing Scientist** to research and develop novel methodologies for determining human cognitive models from multi-lingual sources. Interested candidates should have demonstrated expertise in computational linguistics and natural language processing research as well as experience with software engineering and system integration.

At Two Six Labs you will join a small, multidisciplinary team of researchers from industry and academia that values cooperation and creativity. In this role you will become familiar with critical global health problems and analyze large amounts of language-based data to determine underlying patterns that drive behavior.

Qualifications

- Foundational knowledge of tools and methodologies commonly used to characterize and analyze language (e.g. BERT, word embeddings).
- Experience with Python and common data formats.
- Excellent communication skills.
- Demonstrated experience working as part of a multidisciplinary team.

The successful candidate will have:

- 3+ years work experience and/or a graduate degree in the field of computational linguistics, natural language processing, or related topic.
- Keen understanding of advanced solutions and technical acumen.
- Understanding of common integration methodologies (e.g. containerization, devops).
- Ability to obtain and maintain a DoD clearance.

Nice to Have:

- Knowledge of global health challenges in sub-Saharan Africa.
- Understanding of broader machine learning topics.
- Fluency in an Afroasiatic language.
- Active TS/SCI clearance.

Text Mining Jobs

What about the job that doesn't exist yet?

Text Mining Jobs

Skills Learned

- Basic understanding of using Python
  - Handling dataframes, plotting, opening text files, basic operations
- Normalizing text through pre-processing
  - Find out which words are rare or common with TF-IDF matrix
- Automatically classify text as Good/Bad, Hawkish/Dovish, etc.
  - Naive Bayes Classifier/Support Vector Machines
- Discover topics among a set of unannotated documents
  - Topic Modeling
- Model words in such a way to preserve semantic meanings
  - Word embeddings/Word2Vec
- Thinking above and beyond project expectations...

That’s All Folks

Good Luck with the Final Project and Final Exams!