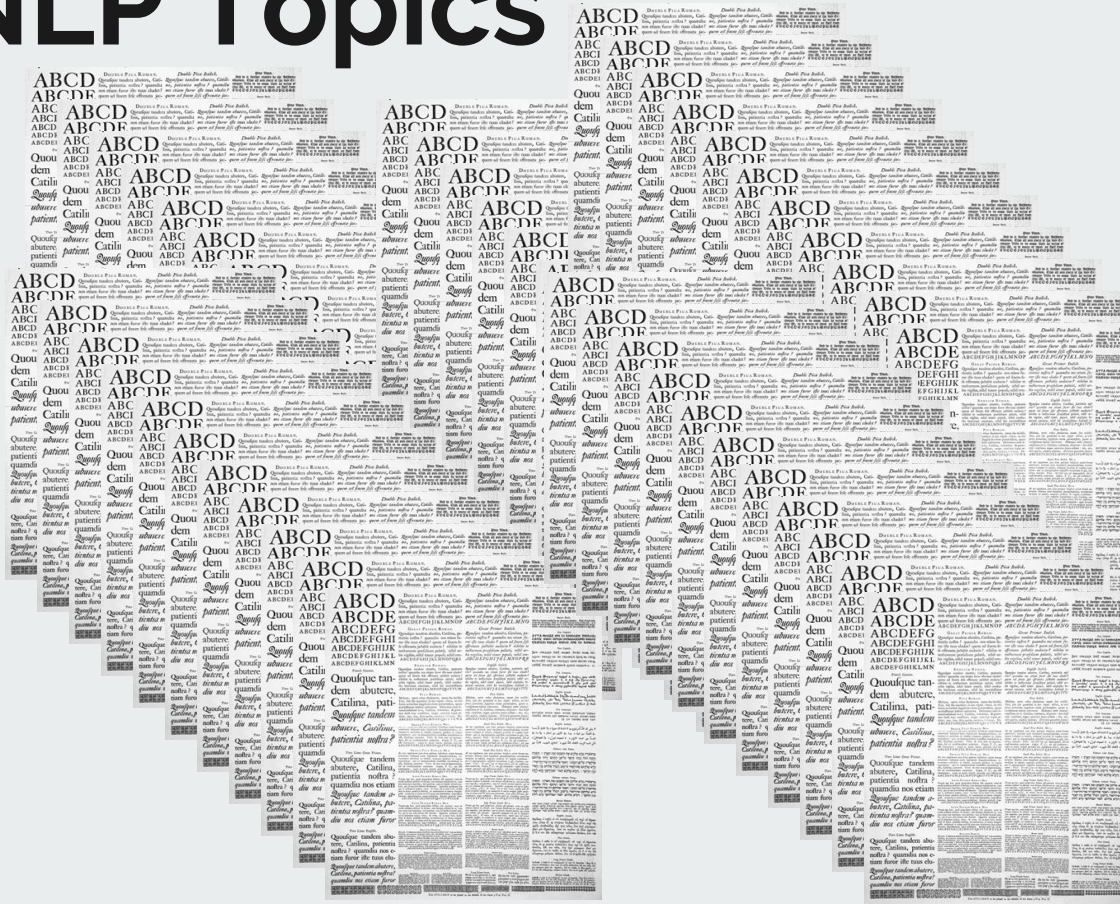
Advanced NLP Topics

Sentiment Analysis, Topic Modeling, and Summarization

Advanced NLP Topics

High Level Problem

Millions of text documents.

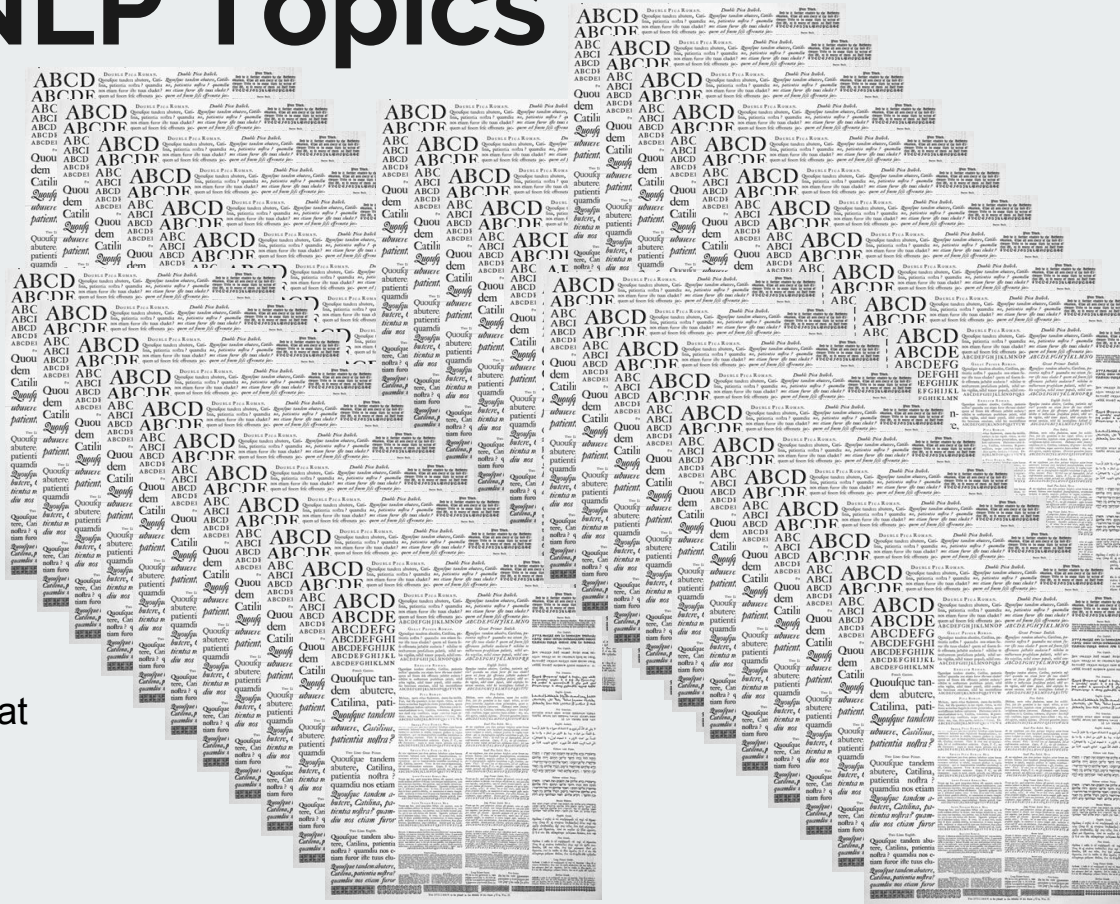


Advanced NLP Topics

High Level Problem

Millions of text documents.

How do we aggregate and understand what the documents contain?



Advanced NLP Topics

Talk Outline

- Summarization
- Topic Modeling
- Sentiment Analysis

- Overview goal of type of analysis
- Common algorithms or approaches
- Python packages for the type of analysis

Advanced NLP Topics

Sentiment Analysis

Quantifies the the subjective ‘emotion’ in a text

Commonly:
Rule based
Also can use:
Knowledge Base
Latent Semantic Analysis (LSA)
Support Vector Machines (SVM)

Advanced NLP Topics

Sentiment Analysis

Quantifies the the subjective ‘emotion’ in a text

Commonly:
Rule based
Also can use:
Knowledge Base
Latent Semantic Analysis (LSA)
Support Vector Machines (SVM)

Topic Modeling

Finds abstract concepts that occur in a body of texts (corpus in NLP)

Commonly:
Latent Dirichlet Allocation (LDA)
Latent Semantic Analysis (LSA)
Also can use:
Non-Negative Matrix Factorization (NMF)

Advanced NLP Topics

Sentiment Analysis

Quantifies the the subjective ‘emotion’ in a text

Commonly:
Rule based
Also can use:
Knowledge Base
Latent Semantic Analysis (LSA)
Support Vector Machines (SVM)

Topic Modeling

Finds abstract concepts that occur in a body of texts (corpus in NLP)

Commonly:
Latent Dirichlet Allocation (LDA)
Latent Semantic Analysis (LSA)
Also can use:
Non-Negative Matrix Factorization (NMF)

Summarization

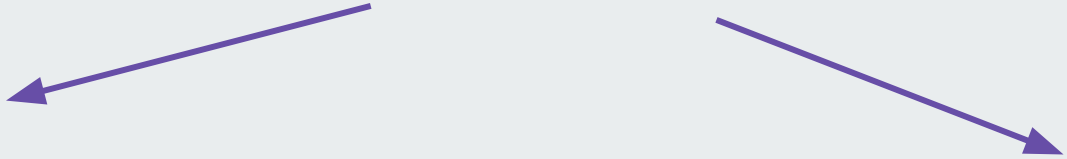
Reduces a text to several key phrases or a representative sentence.

Commonly:
TextRank (Keyphrase)
LexRank (Sentence)
Also can use:
Knowledge Base/ Knowledge graphs

Summarization: Keyphrase

Keyphrase summarization

Finds phrases that represent the document



Extractive summarization

Finds documents out of a collection that summarize what collection is about

Abstractive summarization

Finds documents out of a collection that summarize what collection is about

Summarization: Keyphrase

"The Army Corps of Engineers, rushing to meet President Bush's promise to protect New Orleans by the start of the 2006 hurricane season, installed defective flood-control pumps last year despite warnings from its own expert that the equipment would fail during a storm, according to documents obtained by The Associated Press"

Extractive summarization

- "Army Corps of Engineers"
- "President Bush"
- "New Orleans"
- "defective flood-control pumps"

Abstractive summarization

- Government agency
- Presidential orders
- Defective equipment
- Storm preparation
- Hurricane Katrina

Summarization: Keyphrase

"The Army Corps of Engineers, rushing to meet President Bush's promise to protect New Orleans by the start of the 2006 hurricane season, installed defective flood-control pumps last year despite warnings from its own expert that the equipment would fail during a storm, according to documents obtained by The Associated Press"

Extractive summarization

- "Army Corps of Engineers"
- "President Bush"
- "New Orleans"
- "defective flood-control pumps"

Solved-ish problem: TextRank

Abstractive summarization

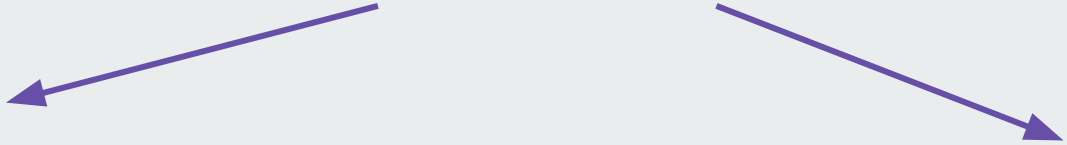
- Government agency
- Presidential orders
- Defective equipment
- Storm preparation
- Hurricane Katrina

Very hard, unsolved: Knowledge Graphs are helpful

Summarization: Sentences

Sentence summarization

Finds a sentence that represent the document



Extractive summarization

Finds the most representative sentence of all the sentences in the document.

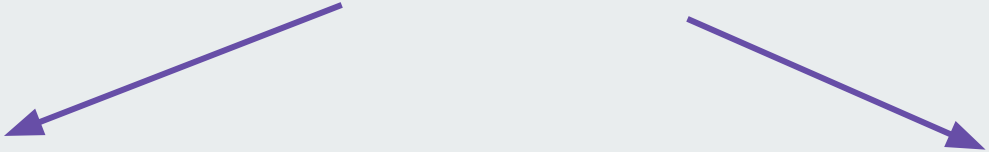
Abstractive summarization

Generates a representative sentence

Summarization: Sentences

Sentence summarization

Finds a sentence that represent the document



Extractive summarization

Finds the most representative sentence of all the sentences in the document.

LexRank

Abstractive summarization

Generates a representative sentence

Very hard, unsolved: Knowledge Graphs plus language generators (auto encoder/decoder LSTMs)

Summarization

Common Packages to Use in Python

Sumy (https://github.com/miso-belica/sumy)

Topic Modeling

Common models attempt to find some topic structure in documents.

Models find the probability that words occur together and then they are grouped into topics.

Topic “Cat”

Milk
Meow
Kitten

Topic “Dog”

Bone
Bark
Puppy

Topic Modeling

Latent Dirichlet Allocation (LDA)

Finds probabilities that words occur together.

Works well on large datasets with longer documents.

Topic Modeling

Latent Dirichlet Allocation (LDA)

Finds probabilities that words occur together.

Works well on large datasets with longer documents.

Latent Semantic Analysis (LSA)

Finds probabilities that words occur together.

“Bayesian LDA”
Works well on small document collections with short documents. Will converge with LDA if document collection is large.

Topic Modeling

Latent Dirichlet Allocation (LDA)

Finds probabilities that words occur together.

Works well on large datasets with longer documents.

Latent Semantic Analysis (LSA)

Finds probabilities that words occur together.

“Bayesian LDA”
Works well on small document collections with short documents. Will converge with LDA if document collection is large.

Non-Negative Matrix Factorization (NMF)

Factors a document-term matrix to find hidden features in documents and clusters them my minimizing an error function

With KL divergence as the error function NMF is the same as LSA.

Topic Modeling

Common Packages to Use in Python

Gensim (https://radimrehurek.com/gensim/)
Textacy (http://textacy.readthedocs.io/en/latest/api_reference.html)
Scikit-Learn (http://scikit-learn.org/stable/modules/decomposition.html#decompositions)

Sentiment Analysis

Reduce documents to the emotion they represent

Sentiment Knowledge bases

Have the obvious affect words (happy, sad, good, bad)
Also, assign probabilities of a word being positive or negative affect to arbitrary words

Sentiment Analysis

Reduce documents to the emotion they represent

Sentiment Knowledge bases

Have the obvious affect words (happy, sad, good, bad)
Also, assign probabilities of a word being positive or negative affect to arbitrary words

Semantic Networks

Uses a vector space to understand more subtle language connected to obvious targets

Sentiment Analysis

Reduce documents to the emotion they represent

Sentiment Knowledge bases

Have the obvious affect words (happy, sad, good, bad)
Also, assign probabilities of a word being positive or negative affect to arbitrary words

Semantic Networks

Uses a vector space to understand more subtle language connected to obvious targets

Rule Based

Uses rules about punctuation, negations, word shape, and modifiers (e.g. “very”, “kind of”)

Sentiment Analysis

Common Packages to Use in Python

NLTK vaderSentiment (https://www.nltk.org/_modules/nltk/sentiment/vader.html)