FROM UNSTRUCTURED TEXT TO STRUCTURED DATA – PART I

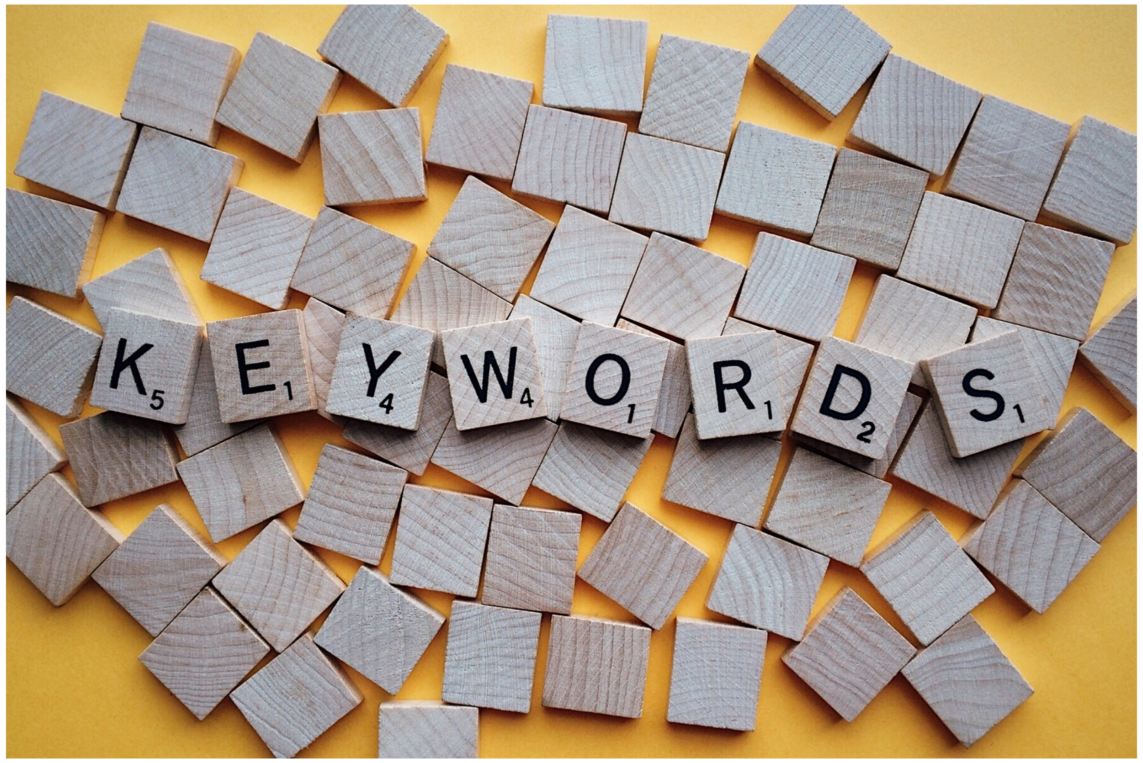
DR. MICHAEL FIRE

Lecture Motivations:

- Learn to analyze massive amounts of text
- Transfer unstructured texts to structured data

Bag-of-Words

A simple text representation model. Using the model, we count the number of times each word appears in a document.



Bag-of-Words

Pros:

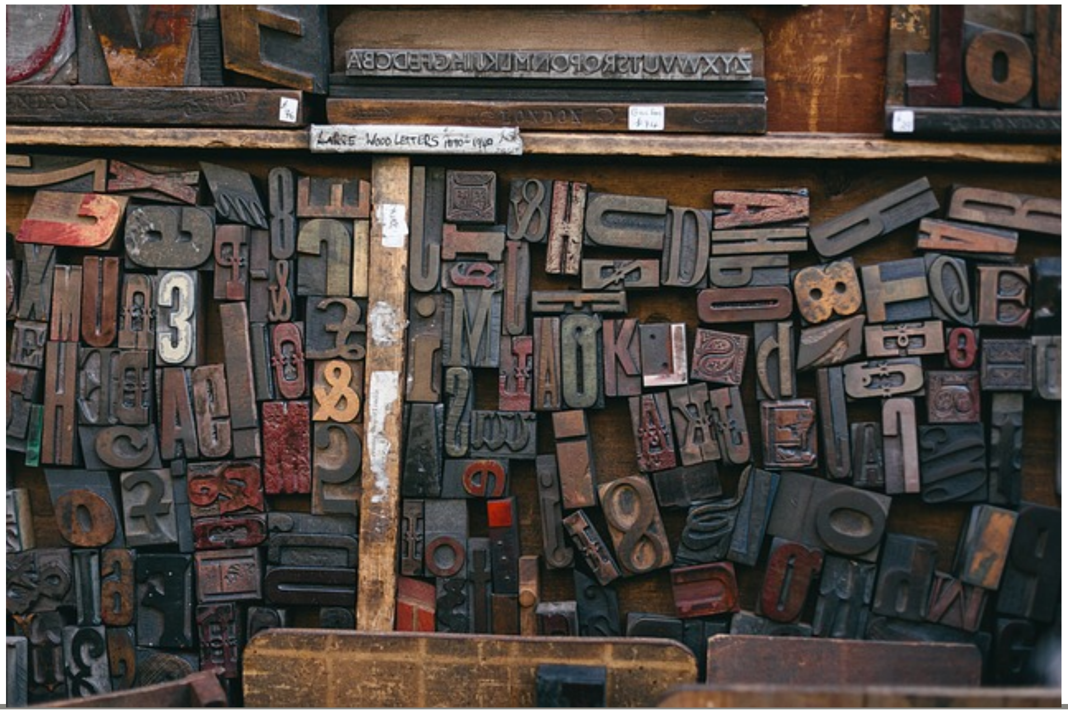
- Easy to understand and simple to use
- Usually provides decent results

Cons:

- Creates many features (tens-of-thousands of features)
- Doesn’t take into account other documents
- Doesn’t provide state-of-the-art results
- It’s possible to miss important characters by removing punctuation

N-Grams

“N-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application” (from Wikipedia)



N-Grams

Pros:

- Easy to understand and simple to use
- Can be used both for “characters” or “words”
- Can utilize useful punctuation and other special characters
- Usually provides decent results

Cons:

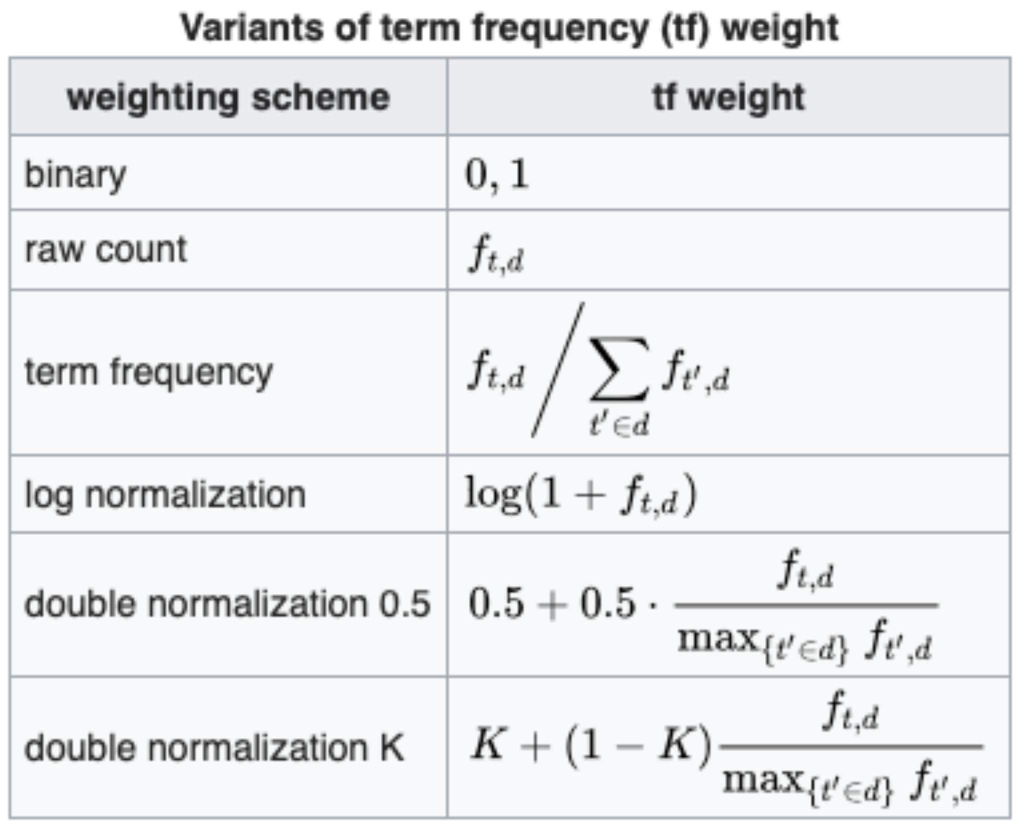
- Creates many features (hundreds-of-thousands of features)
- Doesn’t take into account other documents
- Doesn’t provide state-of-the-art results

Term Frequency–Inverse Document Frequency (TF-IDF)

- TF-IDF is used to reflect **how important a word** is to a **document in a <u>collection</u>**
- A word's TF–IDF value increases proportionally to the number of times it appears in a document and is offset by the number of documents in the corpus that contains it

Term Frequency

A **term frequency**, *tf(t,d), can be the number of time a word appears in a document* (there can also be other measures)



Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

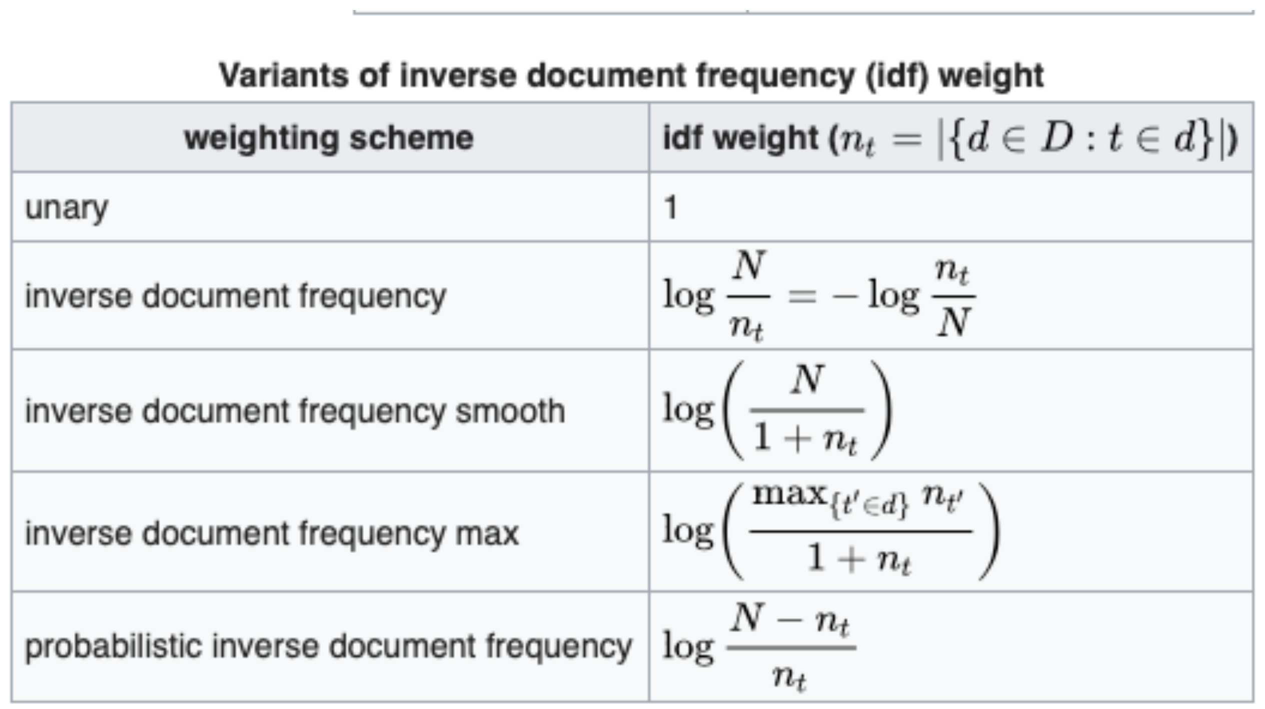
see Wikipedia

Inverse Document Frequency

The **inverse document frequency**, *idf(t,D)*, measures if words are common or rare across all documents.

Usually it is calculated as the logarithm of the number of documents in the collection divided by the number of documents in which the word or term appears.

tfidf(t,d,D) = tf(t,d)*idf(t,D)



Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right)$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N-n_t}{n_t}$

see [Wikipedia]

TF-IDF

Pros:

- Easy to understand and simple to use
- Usually provides decent results
- Takes into account other documents

Cons:

- Creates many features (tens-of-thousands of features)
- Doesn’t provide state-of-the-art results

Topic Model

A topic model is a statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic model algorithms are used to discover hidden subjects in a large collection of unstructured texts

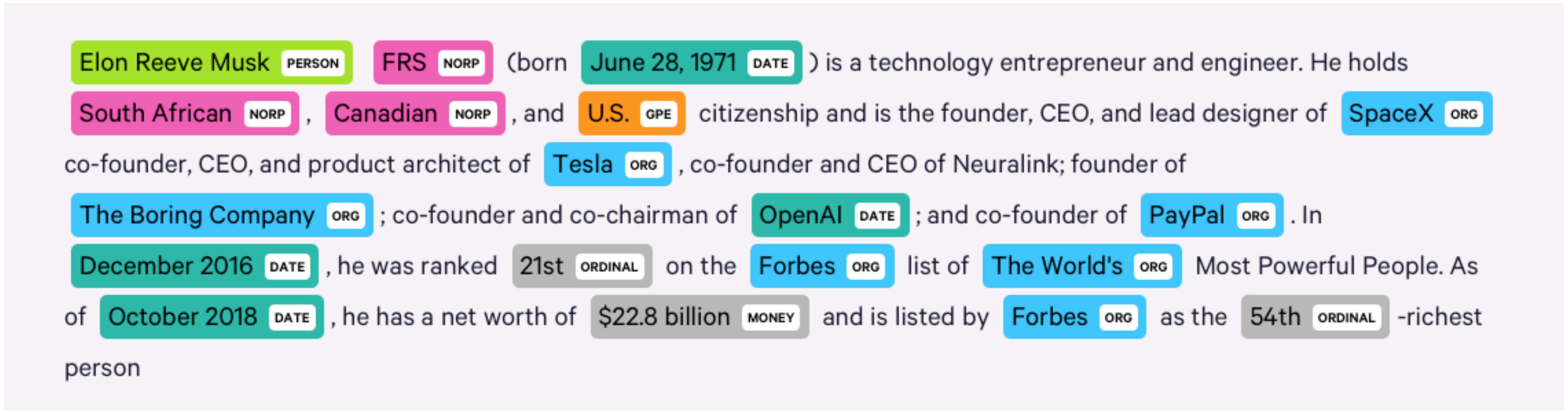
Latent Dirichlet Allocation

"A generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar."

- **A document is a collection of topics**
- **A topic is collection of keywords**

Named-Entity Recognition
Entity Extraction

Entity Extraction is a task that seeks to locate and classify named entity mentions in unstructured text into predefined categories.


Elon Reeve Musk PERSON FRS NORP (born June 28, 1971 DATE) is a technology entrepreneur and engineer. He holds South African NORP , Canadian NORP , and U.S. GPE citizenship and is the founder, CEO, and lead designer of SpaceX ORG co-founder, CEO, and product architect of Tesla ORG , co-founder and CEO of Neuralink; founder of The Boring Company ORG ; co-founder and co-chairman of OpenAI DATE ; and co-founder of PayPal ORG . In December 2016 DATE , he was ranked 21st ORDINAL on the Forbes ORG list of The World's ORG Most Powerful People. As of October 2018 DATE , he has a net worth of $22.8 billion MONEY and is listed by Forbes ORG as the 54th ORDINAL -richest person

Sentiment Analysis

Sentiment analysis (also referred to as opinion mining) is using NLP to identify, extract, quantify, and study affective states and subjective information.

Word Embeddings

“Word embeddings is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers” (Wikipedia)

Word Embeddings

Pros:

- Easy to use
- In many tasks provides state-of-the-art results

Cons:

- Small features space
- Can be domain sensitive (can be missing words/spelling sensitive)

Recommended Read:

- Topic Modeling with Gensim by Selva Prabhakaran
- A Practitioner's Guide to Natural Language Processing (Part I) — Processing & Understanding Text by Dipanjan (DJ) Sarkar
- Word Embeddings in Python with Spacy and Gensim by Shane Lynn
- Word Embedding Using BERT In Python, Anirudh S
- BERT Word Embeddings Tutorial, Chris McCormick and Nick Ryan