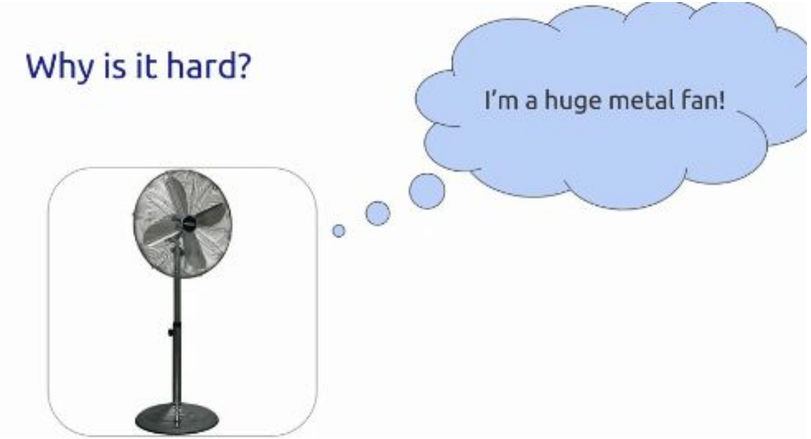
# Natural Language Projects

A survey of topics and recent advancements

Name
Date

// FLATIRON SCHOOL

# Agenda



NLP is hard.

**Problem Spaces**

- Natural Language Understanding
- Natural Language Generation

**Sentiment Analysis**

- Brief history
- VADER

**Topic Modeling**

- Latent Dirchlet Allocation
- Evaluation Metrics

**Embeddings**

- Word2Vec
- Current SOTA

# NLUnderstanding

*Did you just say what I think you just said?*

**Description**

Post-processing of text utilizing context to discern meaning of sentences (sometimes fragmented or run-on) to determine intent

# NLUnderstanding

*Did you just say what I think you just said?*

## Description

Post-processing of text utilizing context to discern meaning of sentences (sometimes fragmented or run-on) to determine intent

**Parts of speech tagging**
Let’s *band* together vs. I want to start a *band*

**Machine Translation**

**Voice Activation**

# NLGeneration

*How do I respond based to what's been said?*

**Summarization**

**Machine Translation**

**Chat Bots**

## Description

In some ways the opposite of NLU:

*Sequence of words <-> General concept*

The choice of a specific, self-consistent representation of a concept which could be expressed in many potential sequences.

# Sentiment Analysis

## AKA Opinion Mining

Seeks to identify and extract a measure of the opinions, attitudes, sentiments or emotions of the writer of the text.

How can we objectively measure something that is subjective?

# *Sentiment Analysis*

## *History*

**LIWC**: Linguistic Inquiry and Word Count

- Hand constructed dictionary of 4500 words, 76 categories, 905 of which in Positive and Negative Emotion
- Internally and externally validated over decades but does not give an intensity of sentiment

**ANEW**: Affective Norms for English Words

- Normative emotional ratings for 1034 words ranked in terms of pleasure, arousal and dominance (score from 1-9)

# *Sentiment Analysis*

## *History*

**SentiWordNet**:

- 147k synsets annotated with 3 scores (positive, negative, neutral) summing to 1
- Very noisy (most synsets are just neutral)

**SenticNet:**

- Publically available semantic and affective resource for concept level opinion and sentiment analysis
- Uses sentic computing, which exploits AI and Semantic Web techniques using graph-mining and dimensionality reduction
- Has a polarity score for concepts like wrath, adoration, woe, and admiration from -1 to 1

# *Sentiment Analysis*

## *VADER*

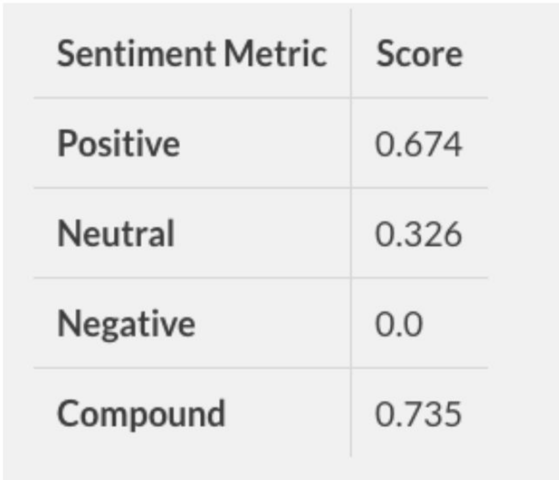
### *(Valence Aware Dictionary for sEntiment Reasoning)*

**Utilizies**

- SentiWordNet for Valence scores based on difference between positive and negative intensity
- SenticNet from an API call

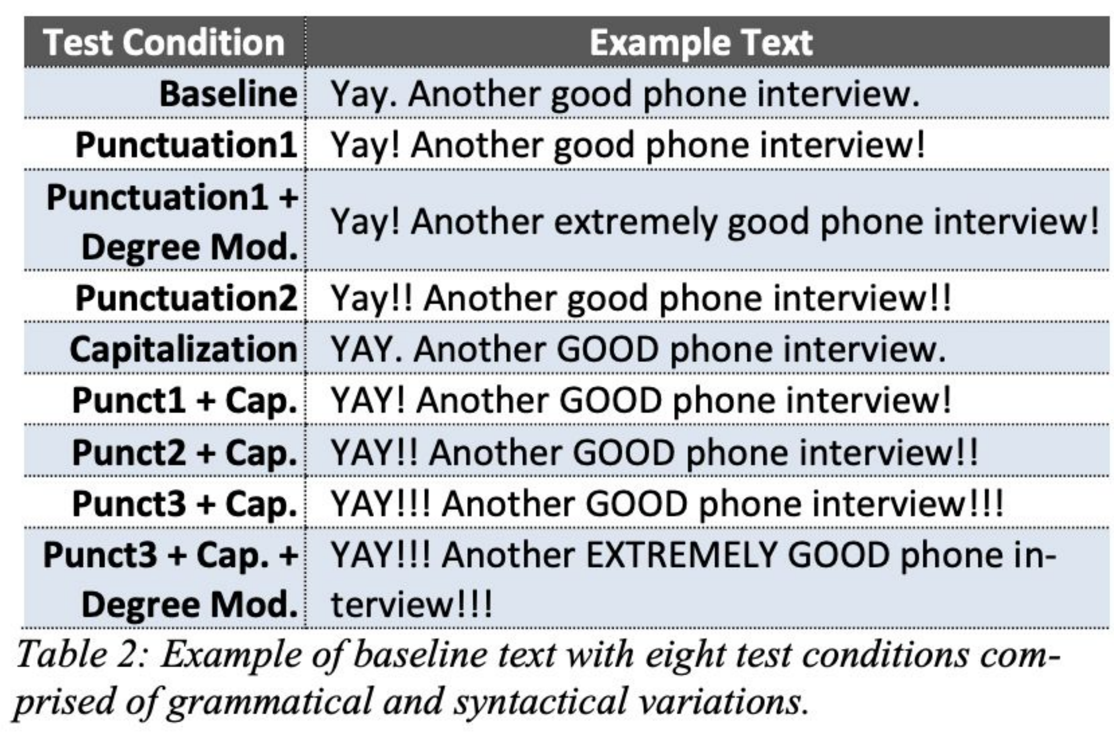
**Lexicon/Context-Awareness**

- Part of speech tagging
- Word sense ambiguation
  - “At first glance the contract looks good, but there’s a *catch*”
  - “The fisherman plans to sell his *catch* at the market.”



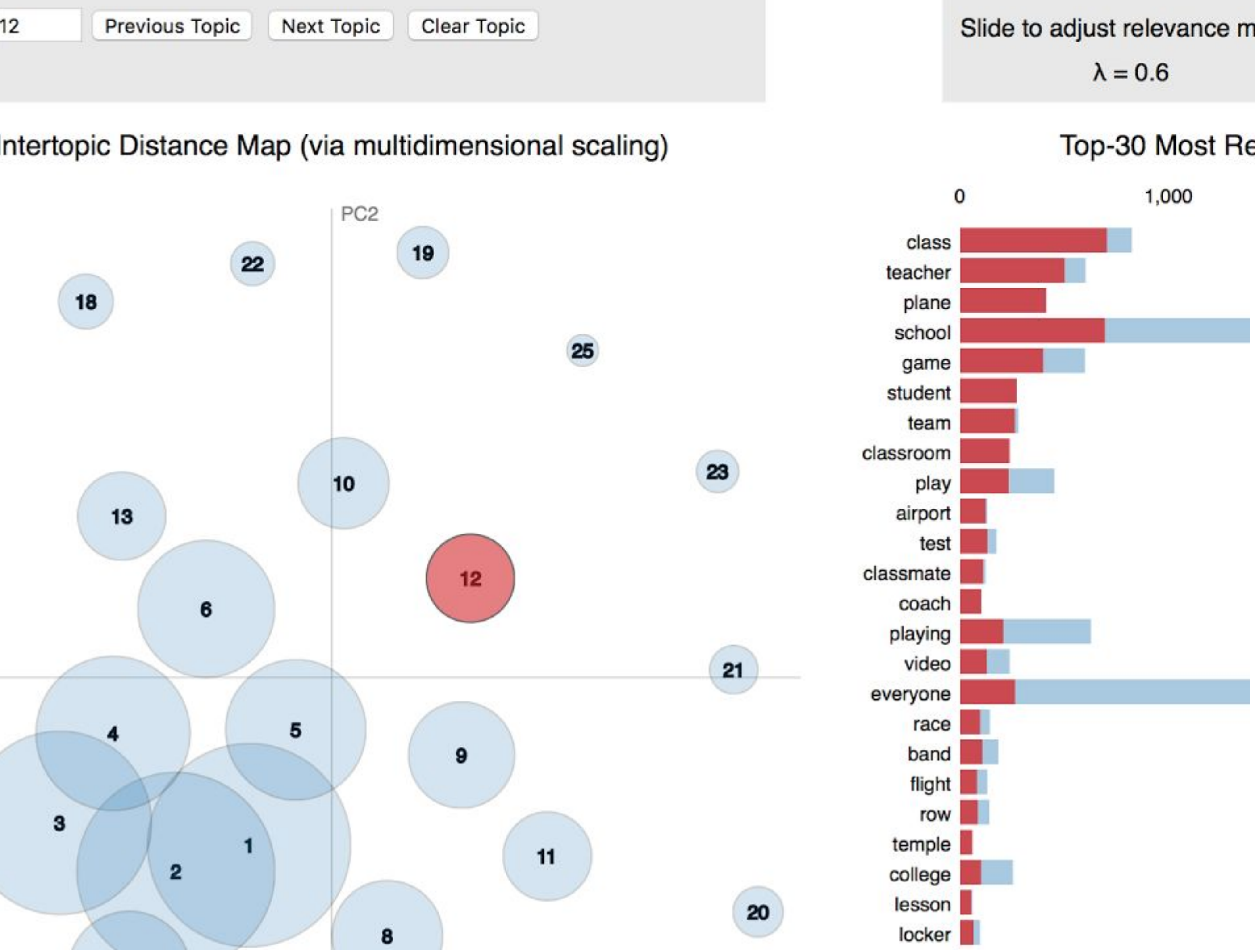
| Sentiment Metric | Score |
|---|---|
| Positive | 0.674 |
| Neutral | 0.326 |
| Negative | 0.0 |
| Compound | 0.735 |

# *Sentiment Analysis*



| Test Condition | Example Text |
|---|---|
| Baseline | Yay. Another good phone interview. |
| Punctuation1 | Yay! Another good phone interview! |
| Punctuation1 + Degree Mod. | Yay! Another extremely good phone interview! |
| Punctuation2 | Yay!! Another good phone interview!! |
| Capitalization | YAY. Another GOOD phone interview. |
| Punct1 + Cap. | YAY! Another GOOD phone interview! |
| Punct2 + Cap. | YAY!! Another GOOD phone interview!! |
| Punct3 + Cap. | YAY!!! Another GOOD phone interview!!! |
| Punct3 + Cap. + Degree Mod. | YAY!!! Another EXTREMELY GOOD phone interview!!! |

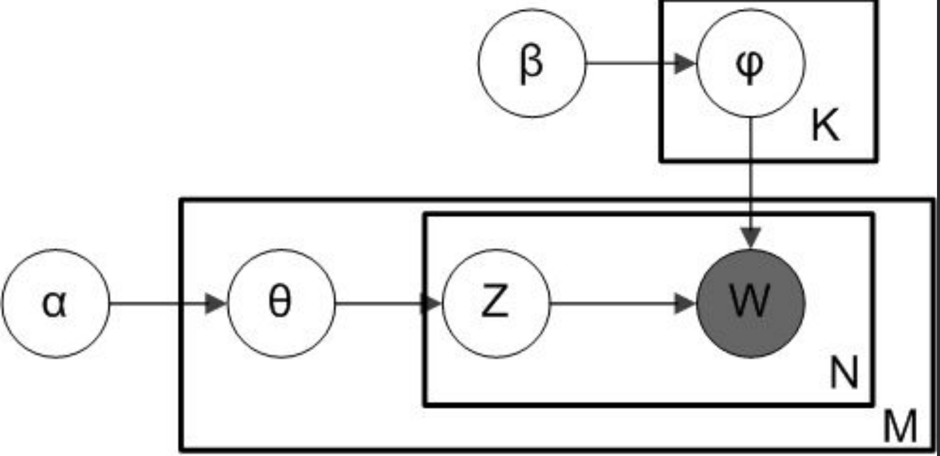
*Table 2: Example of baseline text with eight test conditions comprised of grammatical and syntactical variations.*



# Topic Modeling

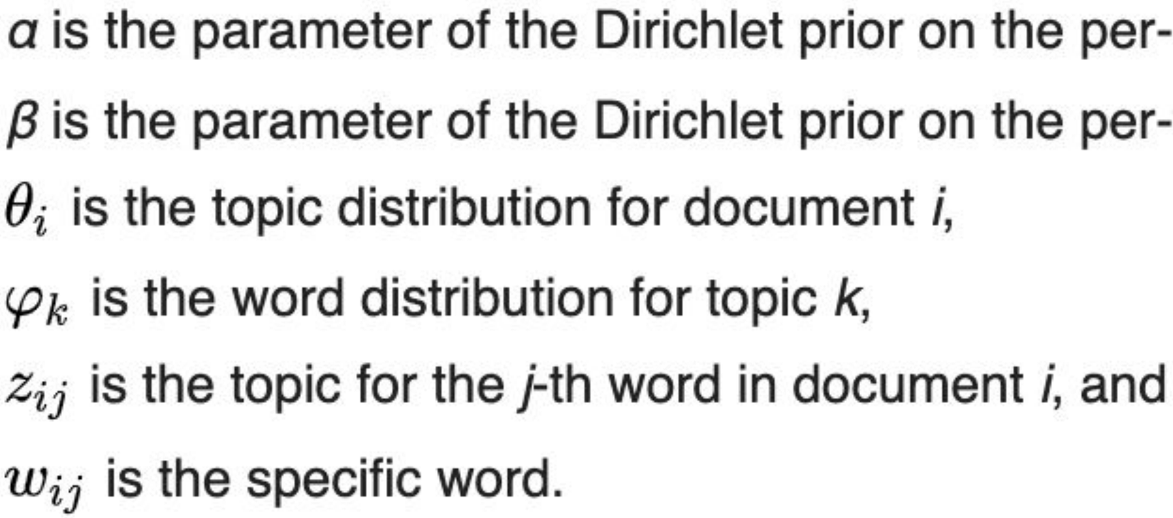
*Documents as a collection of topics*

Find hidden semantic structure in a document by using statistical models to discover abstract “topics” that occur within



# Topic Modeling

*Latent Dirichlet Allocation*



$\alpha$ is the parameter of the Dirichlet prior on the per-document topic distributions,

$\beta$ is the parameter of the Dirichlet prior on the per-topic word distribution,

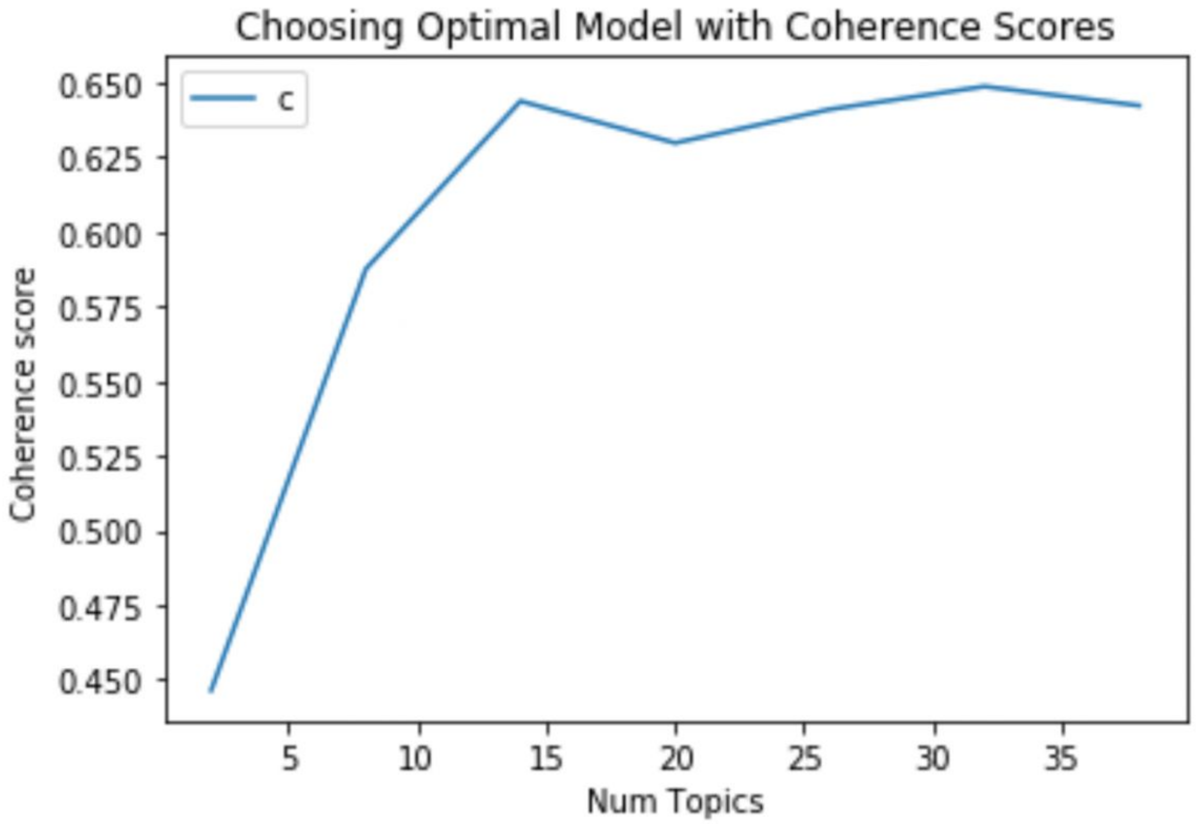
$\theta_i$ is the topic distribution for document *i*,

$\varphi_k$ is the word distribution for topic *k*,

$z_{ij}$ is the topic for the *j*-th word in document *i*, and

$w_{ij}$ is the specific word.

**M** - number of documents
**N** - length of document
**K** - number of topics



# Topic Modeling

*Coherence Scores*

# Embeddings

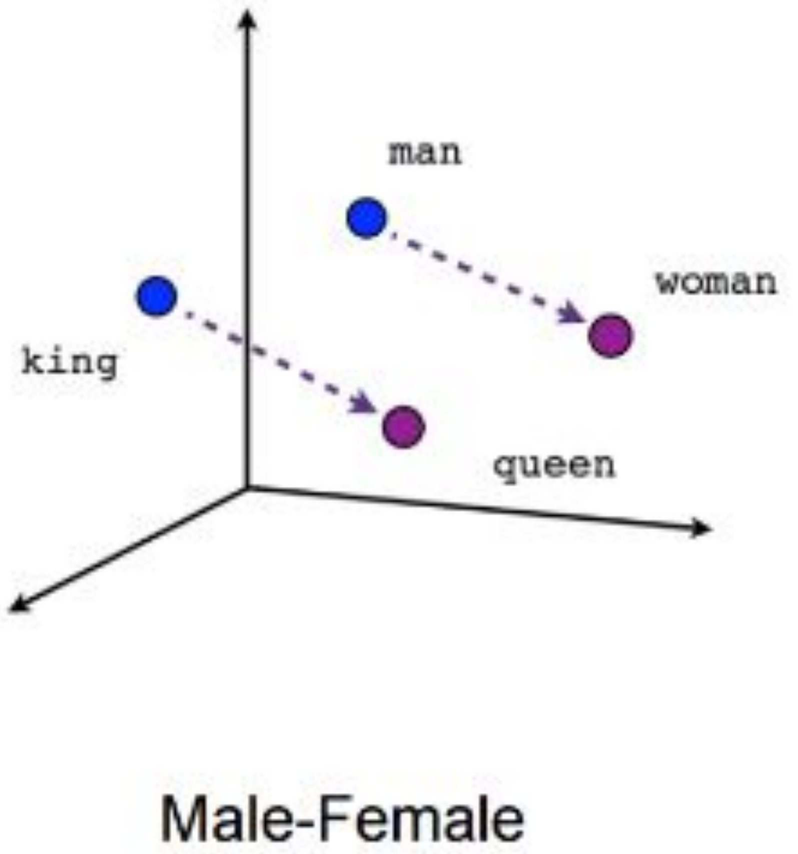
*Meaning in multidimensional space*

## Categorical -> Continuous

An alternative treatment to representing each word as its own feature/token (Bag of Words/Tf-IDF).

- Why not just use One-Hot Encoding?

An embedding is a mapping from a categorical variable to a low-dimensional continuous vector space.



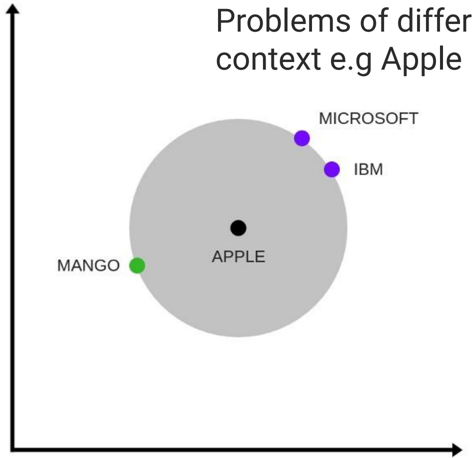
State of the Art

# Embeddings

## Word2Vec

- Continuous Bag of Words
- Skip-gram

Problems of differences in context e.g Apple

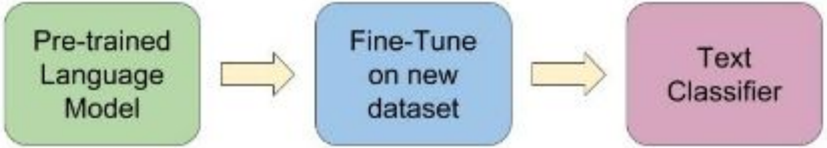


## BERT

- Bidirectional Encoder RepresenTations
- First unsupervised, deeply bidirectional system for pretraining NLP models

## ULMFiT

- Universal Language Model Fine-Tuning
- Transfer Learning from general language model to specific corpus domain



# Resources

### Vader:

- https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f
- http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

### Topic Modeling:

- https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/
- https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

### Embeddings:

- https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526
- https://www.analyticsvidhya.com/blog/2019/03/pretrained-models-get-started-nlp/
- https://bensen.ai/elmo-meet-bert-recent-advances-in-natural-language-embeddings/
- https://www.analyticsvidhya.com/blog/2018/11/tutorial-text-classification-ulmfit-fastai-library/
- https://www.slideshare.net/SebastianRuder/frontiers-of-natural-language-processing