

April 18, 2021

Quinn E Knudsen

IST 736

Syracuse University

What’s the word on AI

uncovering the truth behind the text

**Table of Contents**

**Introduction**……………………………………………………………………………..pg 2

Areas of Focus ………………………………………………………………......pg 2

Artificial Intelligence ...……………………………………………………….....pg 2

Artificial Intelligence Growth ………………………………………………….. pg 3

Machine Learning & Deep Learning………………………………………….….pg 3

Sentiment Analysis ……………..…………………………………………….….pg 4

**Analysis**………………………………………………………………………………..…pg 4

About the Data ...…………………………………………………………….…...pg 4

Analysis 1: Sentences on AI..………………………………………………….... pg 5

Data Cleaning & Prep..……………………………………………………….…..pg 5

Sentiment Classifiers …………..…………………………………………….…..pg 5

Analysis 2: Forbes Articles on Self-Driving Vehicles……………………….…..pg 6

Data Cleaning & Prep……....……………………………………………….….. pg 6

**Results…..**..………………………………………………………………………….…..pg 7

Analysis 1: Sentences on AI……………………………………………….…….pg 7

*Research Question 1………*……………………………………………….…... pg 7 *Research Question 2………*……………………………………………….…... pg 8

Analysis 2: Forbes Articles on Self-Driving Vehicles ..……………………..…..pg 9

*Research Question 3………*………………………………………….….. pg 9

**Conclusions**………………………………………………………………………….…..pg 9

**Follow-up Questions**……………………………………………………………….……pg 10

**Introduction**

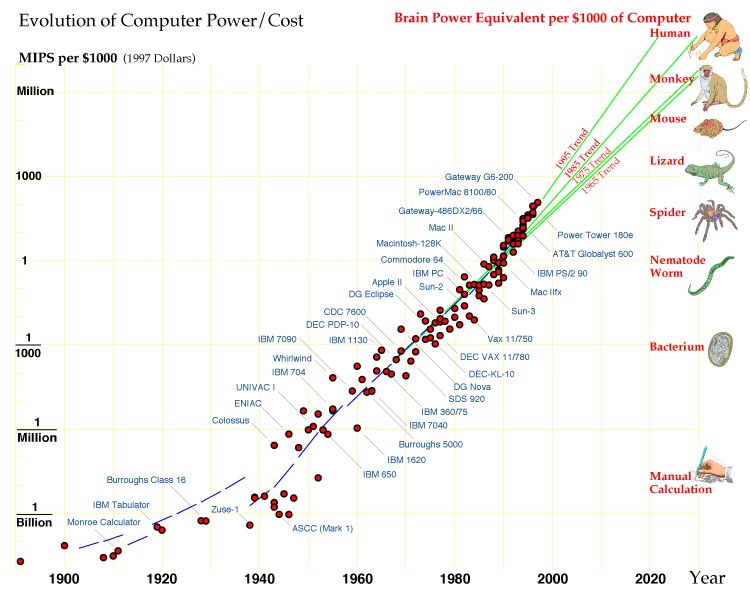
**Areas of Focus**

The scope of this paper is to explore contemporary research on Artificial Intelligence (AI), present a timeline of its maturity, and intersection with Machine Learning (ML). In addition, this paper will explore human perceptions of AI through text analysis drawing from Python packages nltk, Vader, TextBlob, Afinn and two online sentiment analyzers. Finally, this paper will uncover research on Python packages sklearn and CountVectorizer and latent dirichlet allocation (LDA).

**Artificial Intelligence**

Artificial intelligence is the programmatic simulation of human intelligence in machines through a complex process of information retrieval, learning and problem-solving. Although throughout human history there has been a focused prescription of bestowing nonorganic life with human intelligence (i.e. Talos-Ancient Crete 400 BC; Frankenstein 1818) the manifestation of that vision began shifting from fantasy to reality in the 1940s by McCullouch & Pitts 1943 work on artificial neurons. Early research in the 1950s focused on creating AI capable of learning game strategy, solving word problems and proving out logical equations. However, political and computational constraints limited the evolution of AI for several decades.

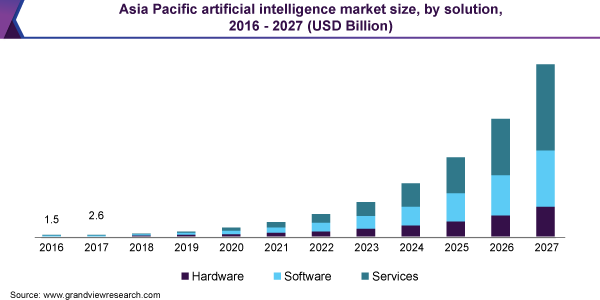
Fast forward to modern times, an AI is not only viewed as an inevitable symbiosis with human existence, the governing bodies of major industrial powers across the world view AI as a strategic advantage for their nation. Part of the changing sentiment on the importance of AI has been the changing computational capability of computers (see Figure 1 below). With personal computers and smartphones today replicating and outpacing the computational capacity of it’s ancestors, the ability to utilize AI as a resource has never been more accessible.



*Figure 1.* Moravec (1998) projects the computational power as an exponential growth curve of computer power relative to cost and capability. <https://www.jetpress.org/volume1/moravec.pdf>

**Artificial Intelligence Growth**

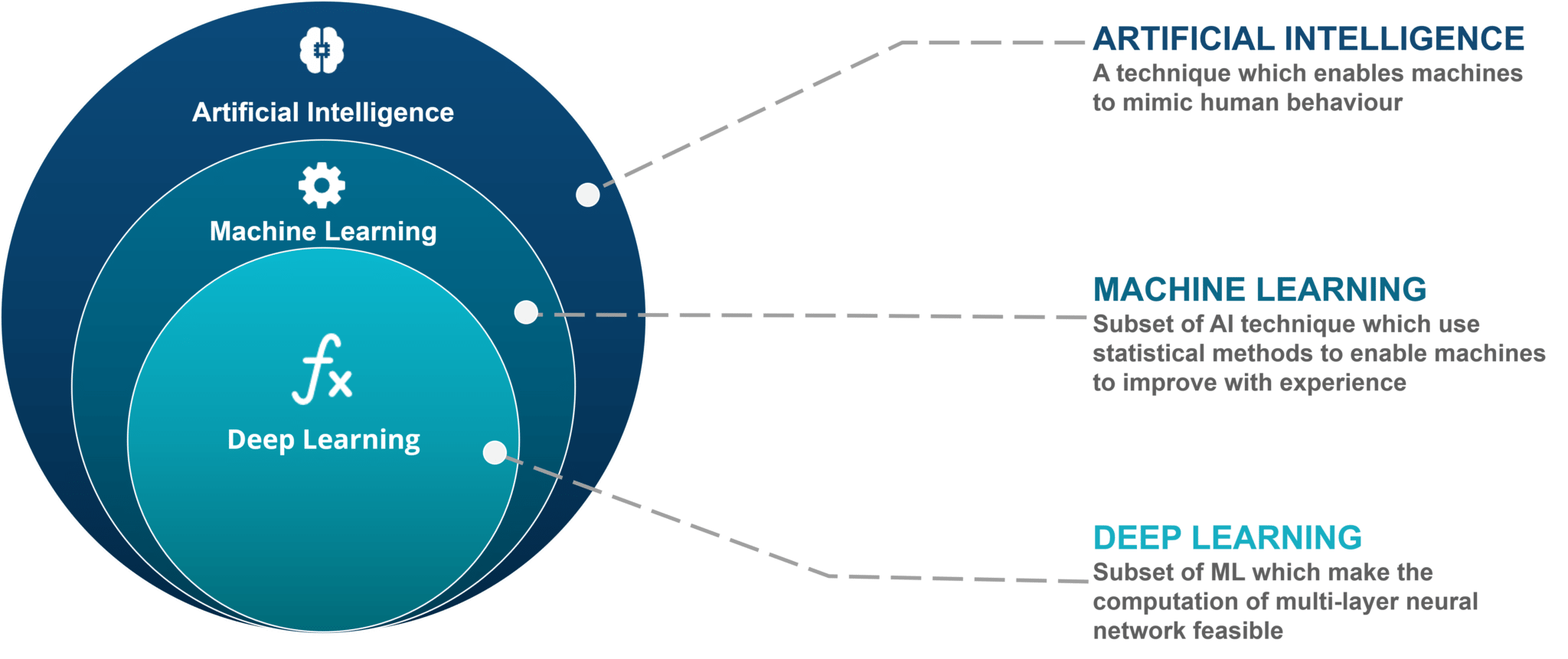
Much like the acceleration of computational power, the acceleration of AI as an industry and the relative speed at which it is becoming embedded into or disrupting established industries has accelerated exponentially in recent years. Statista estimates the global revenue from AI software will grow from $10.1B USD in 2018 to $126B USD in 2025 ([Statista, 2020](https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/)). Grand View Research’s 2020 market analysis of global AI growth expects a compound annual growth rate of the industry of 42.2% from 2020 to 2027 and lead by one of the world’s fastest growing market found in the Asia Pacific (see Figure 2 below) ([Grand View Research, 2020](https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market)).



*Figure 3.* Growth of AI as an industry in Asia Pacific from 2016 to 2027.

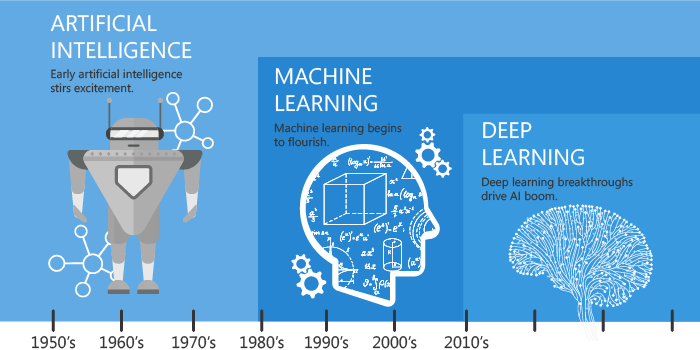
**Machine Learning & Deep Learning**

Machine learning is a facet of artificial intelligence in which complex algorithms and statistical to discern meaning from patterns derived from data. A contemporary approach to hierarchical synergies between the conceptual relationship between AI and ML is illustrated in Figure 3 below.



*Figure 3.* A conceptual map displaying the hierarchies between AI, ML and deep learning <https://top6sites.com/2020/05/15/ai-vs-machine-learning-vs-deep-learning/>

More recently deep learning has flourished with the expansion of big data popularizing algorithms such as neural networks as depicted by Figure 4 below. Built to mirror the manner in which neural activity operates in the human brain, deep learning algorithms leverage a series of nodes, hidden layers and back propagation to predict an outcome variable.



*Figure 4.* Progression of popular AI terminology as an area of focus depicted chronologically. <https://towardsdatascience.com/artificial-intelligence-vs-machine-learning-vs-deep-learning-2210ba8cc4ac>

**Sentiment Analysis**

One applicant of machine learning popularized in recent years is text analytics-or the process of converting large amounts of unstructured written human linguistics into quantitative data capable of uncovering insights meaningful for human interpretation. Sentiment analysis is a specific domain of natural language processing that tries to determine whether the written information is positive, negative or neutral based upon the polarity of the words used in the text. This can prove to be a tall order due to the complexities of language including harder to detect nuances such as sarcasm, slang, negation or amplification. The focus of the analysis portion of this paper will explore an applicant of sentiment analysis on commentary related to AI.

**Analysis**

**About the Data**

Two separate datasets are used to understand the basic nuances of sentiment analysis. The first dataset is a series of written statements about AI to test sentiment efficacy. This dataset contained two columns: a statement about AI and a pseudo-author. 40 statements were written by Notabot Jones while 25 were written by Skylar Net. The second dataset contains two Forbes articles discussing a growing area of AI: self-driving vehicles. One of these articles discussed why some are [opposed](https://www.forbes.com/sites/lanceeliot/2021/04/05/heres-why-some-are-vehemently-and-diametrically-opposed-to-self-driving-cars/?sh=7d135c135699), while the other article brought up reasons why it is [supported](https://www.forbes.com/sites/bernardmarr/2020/07/17/5-ways-self-driving-cars-could-make-our-world-and-our-lives-better/?sh=5176477942a3).

**Analysis 1: Sentences on AI**

**Data Cleaning & Prep**

Two students generated a list of documents in a stream of consciousness related to AI mimicking real world perceptions of the field. As such, the quality of the data was disproportionality clean and required no data cleansing in preparation for sentiment classification.

Data preparation was needed to generate a word cloud. This was achieved through tokenization of the text documents and the removal of common English stop words (from the nltk library) and custom stop words “go” and “going”. Figure 5 below demonstrates the most common words used by the authors with a clear subconscious focus on jobs, data, machine learning, bots, future, and some language indicative of positive sentiment such as love and fun.



*Figure 5.* Depiction of the most common words generated by the authors.

**Sentiment Classifiers**

*TextBlob*

TextBlob uses API for common NLP tasks such as sentiment analysis by assigning polarity (positive, negative or neutral) to a sentence. This classifier was used to compare with the ensuing sentiment classifiers as well as to compare sentiment written by the two authors. Sentiment with a polarity > 0 was considered to be positive, equal to 0 neutral and < 0 negative.

*Afinn*

The AFINN lexicon is a list of English terms manually rated for valence with an integer between -5 (very negative) and +5 (very positive) Årup Nielsen between 2009 and 2011 (<http://corpustext.com/reference/sentiment_afinn.html>).

*Vader*

VADER is a lexicon rule-based sentiment classification tool adept at handling social media and other forms of text. Vader uses a sentiment analyzer to score each text document with sensitivity to polarity (positive or negative) and intensity. While Vader provides scores for negativity, neutrality and positivity, the comparison methodology chosen for this analysis was to compare the compound score. The compound scores each word in the lexicon and then normalized the result between -1 (highly negative) and +1 (highly positive).

*Internet Sentiment Analyzers*

[Free Sentiment Analyzer (danielsoper.com)](https://www.danielsoper.com/sentimentanalysis/default.aspx)

This free tool will allow you to conduct a sentiment analysis on virtually any text written in English. The system computes a sentiment score which reflects the overall sentiment, tone, or emotional feeling of your input text. Sentiment scores range from -100 to +100, where -100 indicates a very negative or serious tone and +100 indicates a very positive or enthusiastic tone.

[MonkeyLearn.com](https://monkeylearn.com/sentiment-analysis-online/)

This free software allows you to quickly drop text into a open field box and with the click of a button it will produce a classification along with a confidence interval. This is a marketing strategy in hopes that you will buy their full services and products.

**Analysis 2: 2 Forbes Articles Discussing AI Applications in Self-Driving Vehicles**

**Data Cleaning & Prep**

Data was scraped from the web and stored in a .csv file for analysis in Python. Data was directly copy and pasted into Daniel Soper’s free web sentiment analyzer.

As before, data preparation was needed to generate a word cloud. This was achieved through tokenization of the text documents and the removal of common English stop words (from the nltk library) and custom stop words “go” and “going”. Figure 6 demonstrates the most common words used in the articles with a clear focus on the topic at hand “self-driving cars”, “AI” and “human” involvement.



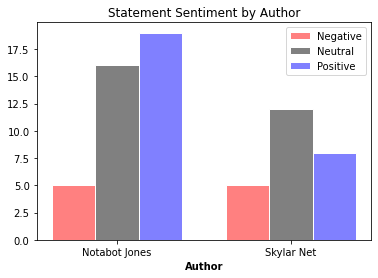
*Figure 6.* Most common words used in the articles on self-driving cars.

**Results**

**Analysis 1: Sentiment by Author**

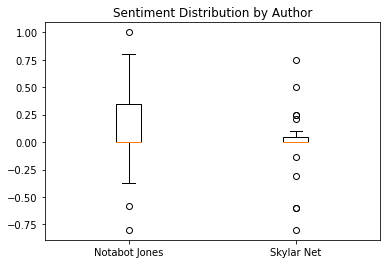
*Research Question 1:* Does author matter in polarity in their statements about AI?

The first step taken was to classify the statements according to their polarity, as visualized in Figure 7, Notabot Jones wrote more tweets, but also appeared to be more favorable in his statements about AI relative to Skylar Net. Both parties were the least likely to be negative when discuss AI according to this classification.



*Figure 7.* Sentiment classification of AI statements reveals general positivity and proportionally more positivity for Notabot Jones.

Next, the range of positivity was explored. As depicted in Figure 8, Skylar Net had many more outliers in her statements and shorter “whiskers”.

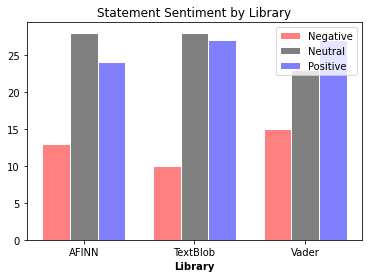


*Figure 8.*  Polarity of statements for the two authors.

To finalize the evaluation of the research question, a t-test was run compare the polarity by author. Results indicated that Notabot Jones (.17) was more positive than Skylar Net (-.01) on average at a 95% confidence threshold.

*Research Question 2:* Do the sentiment analyzers tend to agree on their classifications?

Each sentiment classification was recoded so that positivity was indicated by a value > 0, neutrality = 0 and negativity < 0. For the AFINN, TextBlob and Vader classifications overall positivity, neutrality and negativity were determined. As exemplified in the comparison of Figure 9, negative classification was the least likely categorization. There was variation in that Vader tended to classify the with less neutrality. TextBlob was the least likely to classify the text as negative and like Vader had more positive classifications than AFINN.



*Figure 9.*  Polarity of statements for 3 Python sentiment classifiers for all statements written.

Agreement between classifiers was also tested as depicted in Table 1 below with the highest agreement between AFINN & Vader at 85%. This may partially be explained by the lower negativity found with TextBlob.

|  |  |
| --- | --- |
| **Classifiers** | **Match Rate** |
| **AFINN & TextBlob** | 69% |
| **AFINN & Vader** | **85%** |
| **TextBlob & Vader** | 68% |
| **All** | 63% |

*Table 1.*  Rate of agreement between classifiers finds AFINN and Vader to be the most similar.

**Analysis 2: Sentiment by for Web Articles**

*Research Question 3:* Will the classifiers correctly attribute the stance taken by the authors on self-driving vehicles?

The two articles on self-driving cars were analyzed by the three Python libraries from analysis one and the two free internet sentiment analyzers. As demonstrated in Table 2 below, MonkeyLearn had a hard time making overconfident predictions that the articles were negative, while TextBlob, ever optimistic, found positivity in both.

|  |  |  |
| --- | --- | --- |
| **Classifiers** | **Positive Article** | **Negative Article** |
| AFINN | Positive | Negative |
| Free Sentiment Analyzer | Positive | Negative |
| MonkeyLearn | **Negative** | Negative |
| TextBlob | Positive | **Positive** |
| Vader | Positive | Negative |

*Table 2.*  Rate of agreement between classifiers finds AFINN and Vader to be the most similar.

**Conclusion**

Artificial intelligence is growing and along with it the multitude of applications. As AI infiltrates many fields and industries practitioners and researches find was to apply technology to analyze a variety of data-including text documents. This paper found that two authors subconsciously indicated their feelings towards AI by writing a number of statements on the topic. Additionally, word clouds can be generated to visualize the most common text found in a paper.

Further findings suggest that within the practice of analyzing text for sentiment, there is variable agreement amongst the classifiers. Depending upon the Python package or even free software available on the internet, you may have different interpretations of that data. As research is an iterative process, nonetheless improvements are being made each day in order to handle and help computers make sense of the complexities of human linguistics.

**Follow-up Questions**

2) Next, I would like everyone to research sklearn and CountVectorizer. You do not need to code – we will get to that starting in Week 2. For now – explore, discuss, write about them. What is the difference between input=”content” and input=”filename” for example?

*Sklearn*

This is found within scikit-learn, one of the most useful and widely used Python packages for machine learning. This module integrates classical machine learning algorithms and aims to provide simple and efficient solutions. <https://www.kite.com/python/docs/sklearn>

Within sklearn you can do classification tasks (SVM, nearest neighbor, random forest etc.), regression (SVR, nearest neighbor, random forest etc.), clustering (k-Means, spectral clustering, mean-shift etc.), dimensionality reduction (k-Means, feature selection, non-negative matrix factorization etc.), model selection (grid search, cross validation, metrics etc.), and preprocessing (<https://scikit-learn.org/stable/>).

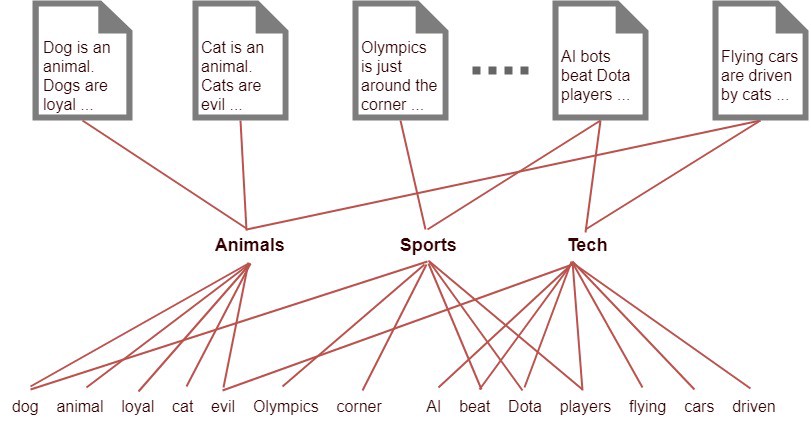
*CountVectorizer*

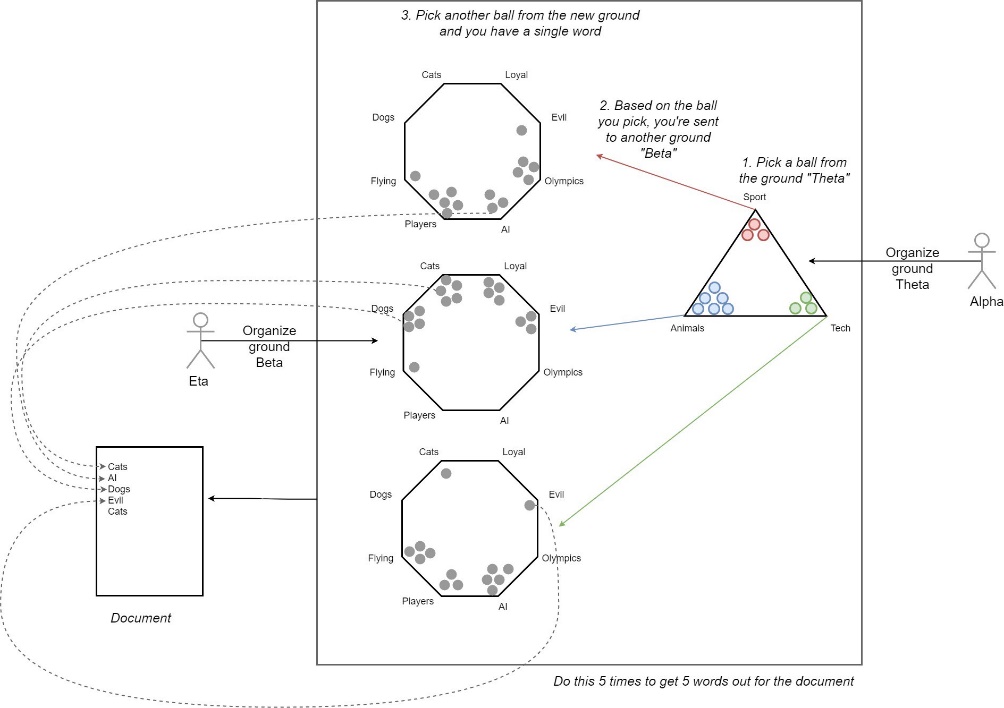
This converts a collection of text documents into a matrix of token counts. For each word in a text document, a numeric value is created. For example, in the sentence, “I love learning” a count of 1 would be recorded for each word in the sentence. If the a second sentence were to follow reading “I love learning about data because I love data” the word “I”, “love”, and “data” would have a 2 and the word because would have a 0 associated with prior sentence. The data becomes high dimensional as the vocabulary increases with each sentence. This is a value tool for preprocessing text data for analysis. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>

Input=”filename” is used when taking a list of files as input for the vectorization. On the other hand, input=”content” means it would pass the actual documents as a list of strings.

3) Finally, read about latent dirichlet allocation (LDA) – it is for topic modeling… Here again, there is no need to code or to use Python. Just start to think about these ideas.

Latent dirichlet allocation (LDA) is a statistical model used for topic modeling. Topic modeling is used to uncover hidden themes in the text, classifying the text and using the classifications to organize or summarize the documents. The general assumptions of LDA are that each document is a collection or “bag of words” without consideration for grammar. Common words (stop words) do not bring value to the topic modeling and are excluded. K-the number of topics we want modeled is known beforehand and follows to logic that certain words have a higher probability of being assigned to the topic. This is a highly useful technique for understand themes in large amounts of text data. For example, modeling the topics in an employee feedback survey.





**Definitions and Notations**

1. k — Number of topics a document belongs to (a fixed number)
2. V — Size of the vocabulary
3. M — Number of documents
4. N — Number of words in each document
5. w — A word in a document. This is represented as a one hot encoded vector of size V (i.e. V — vocabulary size)
6. w (bold w): represents a document (i.e. vector of “w”s) of N words
7. D — Corpus, a collection of M documents
8. z — A topic from a set of k topics. A topic is a distribution words. For example it might be, Animal = (0.3 Cats, 0.4 Dogs, 0 AI, 0.2 Loyal, 0.1 Evil)

[Article 1](https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2) [Article 2 (Super Awesome)](https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158)