News title analyzer

Utilizing machine learning to categorize news titles

Group 18
Martin Ryberg Laude
Saga Jonasson
Markus Brewitz
Vilhelm Norström

The Dataset

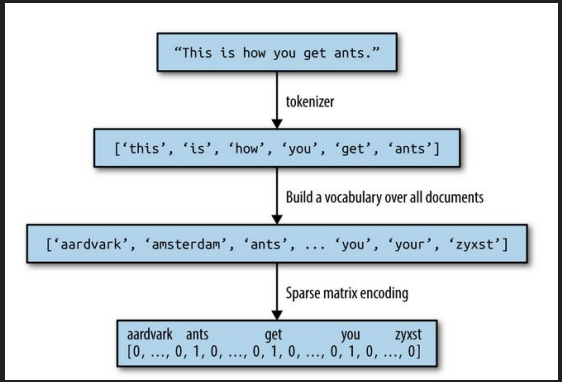
- 400 000+ news articles
- 4 different categories
- Had to reduce for runtime

Make a program that can categorize news articles

Source of the data: Artificial Intelligence Lab @ Faculty of Engineering, Roma Tre University - Italy

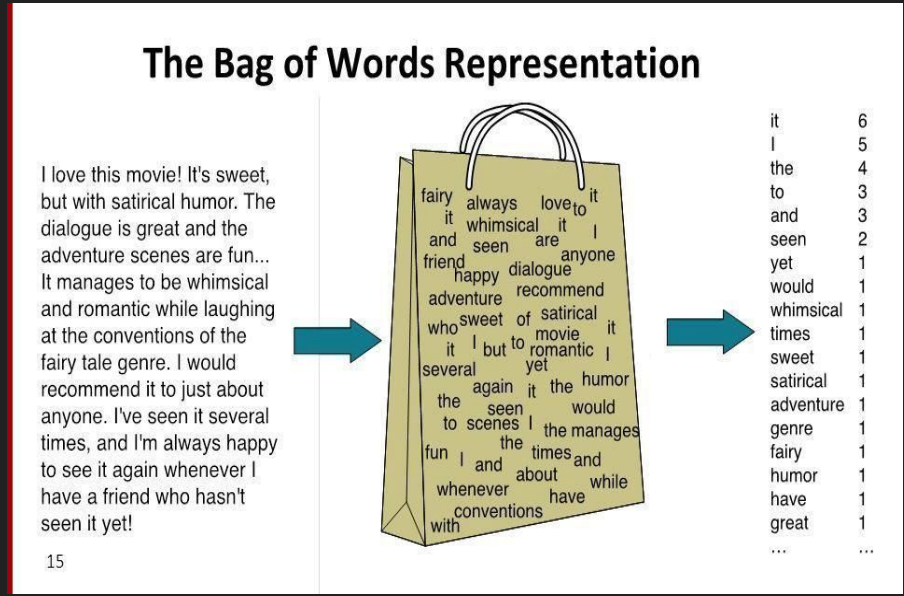
Vectorizing our data

- Bag of Words (BoW)



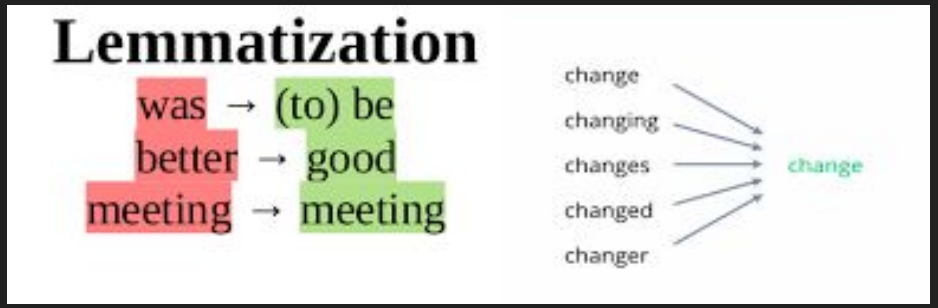
- And TF-IDF (Term Frequency-Inverse Document Frequency)

$$\text{tfidf}(w, d) = \text{tf} * \log\left(\frac{N+1}{N_w+1}\right) + 1$$

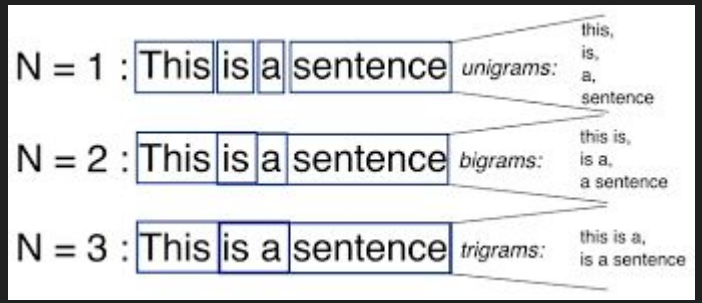


Vectorizing our data

- Lemmatization - grouping inflections of the same word



- N-grams - Taking the order of words into account

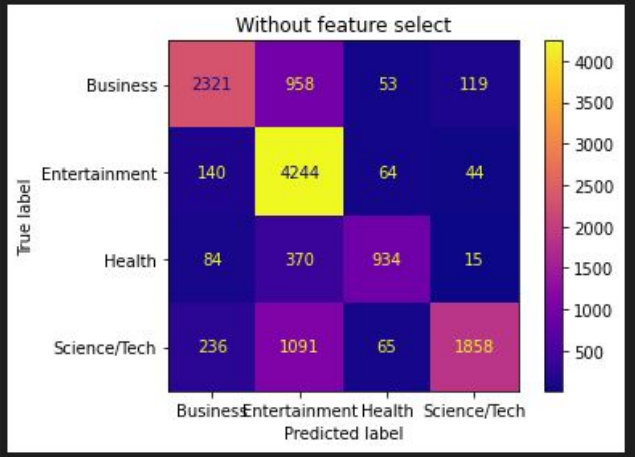


“Not good, it’s bad”

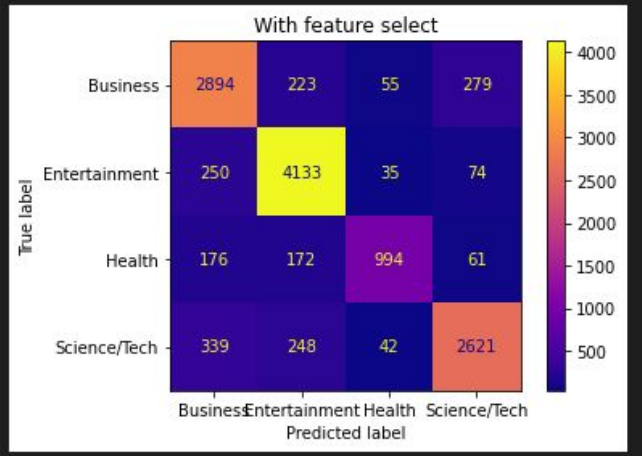
“It’s good, not bad”

Supervised learning

- k-NN
- Naive Bayes
- SVM (Support-Vector Machine) ★
- Feature selection



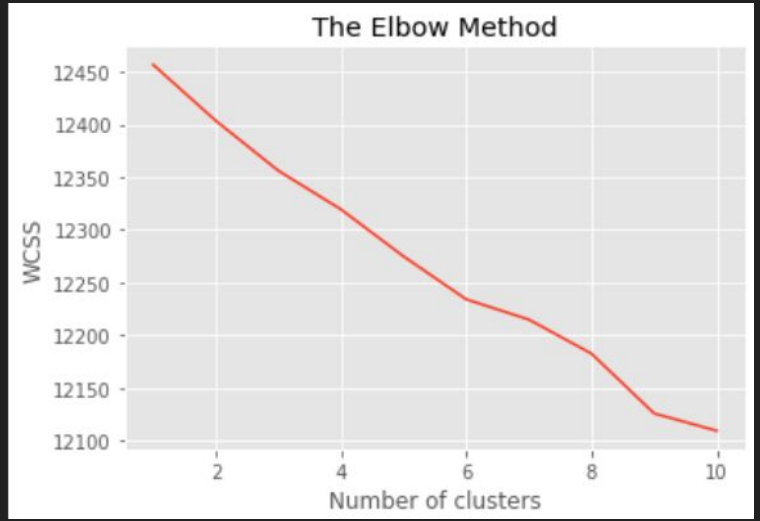
Accuracy: 0.7429

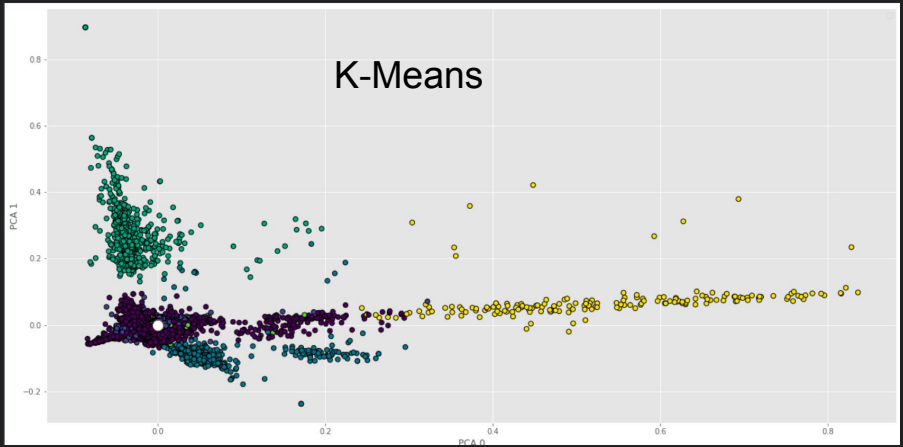


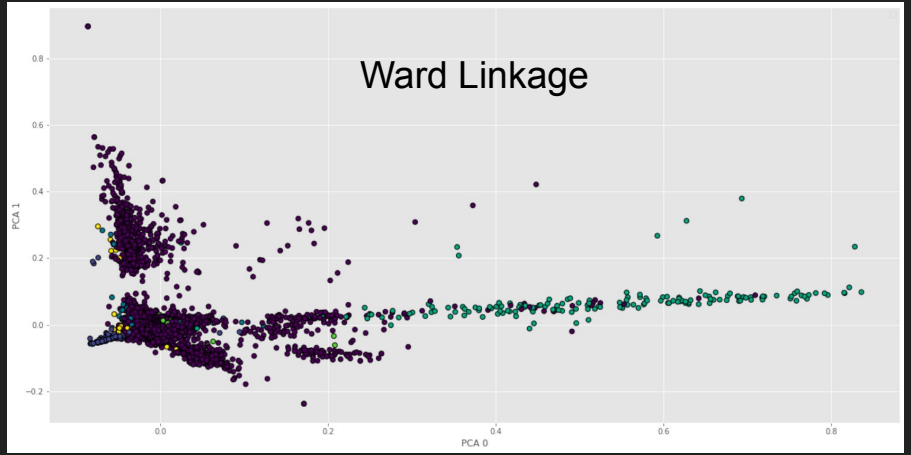
Accuracy: 0.8449

Unsupervised learning

- K-Means
- Elbow method
- PCA
- Comparison: Ward Linkage







```
----Silhuette Coefficient: higher is better----
K-Means: 0.010381475729508 39
Ward: 0.0071764867491022405

----Calinski-Harabasz Index : higher is better----
K-Means: 47.3840547312038
Ward: 38.56540302618462

----Davies-Bouldin Index : lower is better----
K-Means: 6.566354605122719
Ward: 3.3376678912502356
```

```
0 : say, video, year, microsoft, report, watch, day, facebook, time, twitter, review, star, study, price, amazon, season, deal, million, stock, ha, 'the, game, rate, news, tv
1 : new, new york, york, trailer, video, album, release, feature, new album, report, apple, film, google, new film, facebook, season, release new, time, unveils, look, photo, movie, star, high, ha
2 : google, apple, samsung, galaxy, s, samsung galaxy, galaxy s, glass, google glass, beat, android, iphone, buy, launch, note, galaxy note, sale, patent, v, o, feature, tab, price, smartphone, day
3 : u, u s, s, u stock, stock, rate, sale, home, market, china, data, dollar, job, growth, oil, consumer, price, economy, case, say, percent, month, u economy, gain, expected
4 : woman, risk, study, cancer, increase, alzheimer's, disease, thor, death, pelvic exam, exam, pelvic, say, healthy, double, new, inflation, report, heart, suicide, men, lower, reduce, d, train
5 : kim, kardashian, kim kardashian, kanye, wedding, jay z, z, jay, kanye west, west, beyonce, kardashian kanye, kim kardashian's, kardashian's, solange, beyonce jay, rob, beyonc, rob kardashian, west's, tour, dress, vogue, cover, bikini
```

What could be improved?

Optimization!!!