# Question 3 Vector space Model and feature representation

Experiment with different representation techniques. Document your findings and make conclusions. Show how choosing n-gram features can influence your results

1. Representation techniques used
2. Findings (results)
3. Using n-grams (1, 2, 3, 4)
4. How n-grams influence results

For this segment of the course work we aimed to experiment with **count-based** representations. For the count-based representations, we utilized **term frequency** and **inverse-document term frequency** to evaluate the relative importance of each word or feature.

We represented each review as a matrix of **token counts** through the process called Count Vectorization.

<https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>

term frequency is simply accounting for the number of times a term (a word) occurs in a given corpus (collection of documents) it does not differentiate between the context of the term or the document it is located in. A matrix of term frequencies is returned by the algorithm.

tf(*t*,*d*) = *ft*,*d*

The relative weight of each term using turn frequency is calculated by dividing the term frequency by the cumulative frequency to gain the relative frequency of each term.

In contrast inverse document frequency inverts the frequency of the term in the documents Under the assumption that the most common words would be the least indicative of the documents overall meaning.

Idff=log10(N/dff)

in combination turn frequency inverse document frequency indicates the relative weight or importance of a term regarding the entire corpus of documents. This is crucial in feature extraction as it identifies the most important terms in the corpus that can provide insight into the documents overall meaning.

Tf-idf= tf(*t*,*d*)\* Idff

Findings - tff

A picture containing indoor, computer

Description automatically generated

Appendix X shows a sample of the output of the term frequency matrix representation of the data. Each row represents a document in the corpus in our case a review with each column representing a term and its frequency.

Tf-idf

A screenshot of a cell phone

Description automatically generated

Appendix X shows a sample of the output of the TF-IDF matrix representation of the data. In review we conclude that in the context of food reviews TF-IDF is a more informative vector space representation then turn frequency as it indicates the relative value of each word regarding our corpus of reviews.

Furthermore, we experimented with n-grams. By default, unigrams were used to represent features then we attempted to try bigrams and trigrams and quad grams.

Term Frequency – Bigrams

TF-IDF – bigrams

Term Frequency – Trigrams

TF-IDF – Trigrams

Term Frequency – Tera-grams

TF-IDF – Tera-grams

As seen from the above examples, n-grams affected …..