# Word net:

WordNet is a lexical database for the English language

It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records the various semantic relations between these synonym sets.

WordNet is often used for tasks such as text classification, semantic parsing, and word sense disambiguation, as well as for creating new word-sense disambiguation algorithms.

Synset, not the word, is the fundamental unit of WordNet.

## Part

MultiWordNet for differernt language

**Synonymy** : synonyms words

**Hypernymy** : refers to the relationship between a general term (a hypernym) and a specific instance of that term (a hyponym)

**Hpernym** is a more general category,

**Hyponym** is a specific member of any category.

• Hypernym: Vehicle

• Hyponyms: Car, motorcycle, truck, boat

**Holonym** is a word that describes a whole object or concept.

e.g : "body" is a holonym of "arm"

Holonyms can be obtained using the member\_holonyms() or part\_holonyms() or substance\_holonyms() methods

Semcor?

path\_similarity

vs wu\_similarity

# WUP

Wu-Palmer Similarity (WUP Similarity) is a measure of semantic similarity between two synsets (sets of synonyms) in WordNet.

The similarity score is calculated as the depth of the least common hypernym (lch) divided by the sum of the depths of the individual synsets.

WUP Similarity(synset1, synset2) = 2 \* depth(lch(synset1, synset2)) / (depth(synset1) + depth(synset2))

* synset1 and synset2 are the two synsets being compared
* lch(synset1, synset2) is the least common hypernym of synset1 and synset2
* depth(synset) is the depth of synset in the WordNet hierarchy, where the root node (the most general term) has depth 0 and all other nodes have depth n+1 where n is the depth of the parent node

Its range 0 to 1.

 Senseval corpus

Sysnets 🡪 defination, examples, name

# WSD (Word Sense Disambiguation)

All documents and article contains the most frequent word so it is difficult to beat that word.

In nltk corpus : semcor part of brown corpus is used for the sense

For the supervised data ww can use the annotated data whereas for umsupervised data we use lesk algorithm we can say this"distantly-supervised”.

## Lesk algorithm

It is used to compare the definitions of different senses of a word in a corpus, such as WordNet.

* It is easy to use
* It relies on the defination on the different senses.
* On the poor dataset is difficult to find out the actual difference.

Lesk return sysnets object like :

synset = lesk([word] 'interest', pos[0].lower())

## Chi squared(content analysis)

It is used to determine the association between two different lexicons.

* Chi squared is used to determine the reation between two different lexicons like words or phrases from the two lexicons.
* The Chi squared test depends on the count of the data, for better results we need large enough.
* We can perform with Chi squared test with chi2\_contingency function(scipy library). It construct the table and provides a p-value, which is used to check significance of the association.
* Does not tell about directionality of the relation, to get his we have to apply original ratios of the data.
* It is impo to apply multiple hypotheses, the normal p < 0.05 cutoff for statistical significance is no longer valid. To get p value we have apply the Bonferroni correction, which adjusts the p-value cutoff based on the number of hypotheses being tested.
* One popular lexicon for content analysis is the Linguistic Inquiry and Word Count (LIWC, pronounced Luke) . It is paid.
* The Chi squared test is a non-parametric test, which means it does not make any assumptions distribution of the data. This makes it a useful tool for analyzing data that may not fit a normal distribution.
* One limitation of the Chi squared test is that it only considers two-way relationships. If you want to analyze the relationship between more than two variables, we need to perform different statistical test, such as logistic regression or multiple regression.
* The Chi squared test is sensitive to the independence assumption between the two lexicons. With high degree of correlation between the words in the two lexicons, the results of the test may be biased.
* Another limitation of the Chi squared test is that it cannot handle missing data.
* Need large and representative sample of texts with pre-process the data like removing stop words, stemming or lemmatizing the words, and removing outliers or irrelevant data.
* It can be used with sentiment analysis, topic modeling, or network analysis to gain a deeper understanding of the relationship between the two lexicons and the underlying meaning of the text.