**Corpus and WordNet**

1. Introduction:

Ngôn ngữ tự nhiên là ngôn ngữ mà các loài động vật sáng tạo ra để giao tiếp với đồng loại. Ngôn ngữ tự nhiên có 2 dạng là chữ viết và âm thanh (tức tiếng nói).

Corpus là những tài nguyên ngôn ngữ vô cùng căn bản và cần thiết cho NLP. Từ điển là tập hợp các tri thức về ngôn ngữ, cách sử dụng và ý nghĩa của từ, thì ngược lại, corpus là dữ liệu về cách sử dụng, cách viết của từ đó trong thực tế.

Corpus được sử dụng trong thống kê ngôn ngữ học, để quan sát và phân tích các hiện tượng ngôn ngữ.

WordNet là dự án từ điển đơn ngữ (tiếng Anh) thiên về ngữ nghĩa. WordNet đã tạo ra một tập hợp từ vựng đồ sộ, theo đó các từ được sắp xếp trong dãy của những tập hợp đồng nghĩa, giúp cho việc xác định nghĩa của từ và để phân biệt được nghĩa đang xét với các nghĩa khác. Nguyên lí tổ chức chung của WordNet là mạng lưới quan hệ ngữ nghĩa.

Corpora là các dữ liệu văn bản, ngôn ngữ đã được số hoá. Có thể xem Corpora như một kho ngữ liệu. Corpora thường là các dữ liệu đã được xử lý, được sử dụng như đầu vào của các thuật toán trong NLP.

**Các Corpora có sẵn của NLTK.**

* 1. Gutenberg Corpus:

Gutenberg là một dự án cung cấp 25.000 cuốn sách điện tử miễn phí, trang Web của nó là <https://www.gutenberg.org/>. Và NLTK đã lấy 1 phần nhỏ số sách trong dự án này để cung cấp cho chúng ta trong các gói của mình.

* 1. Web và Chat Text Corpus:

Cung cấp các văn bản mang tính trang trọng, các đoạn hội thoại trên Web hay mạng xã hội.

* 1. Brown Corpus:

Đây là Corpus điện tử đầu tiên của tiếng Anh, có từ năm 1961. Nó chứa văn bản từ 500 nguồn và các nguồn thì được chia mục theo thể loại như tin tức, các bài xã luận...

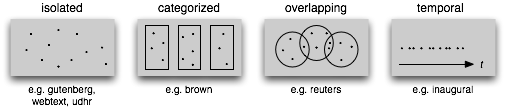
* 1. Reuters Corpus:

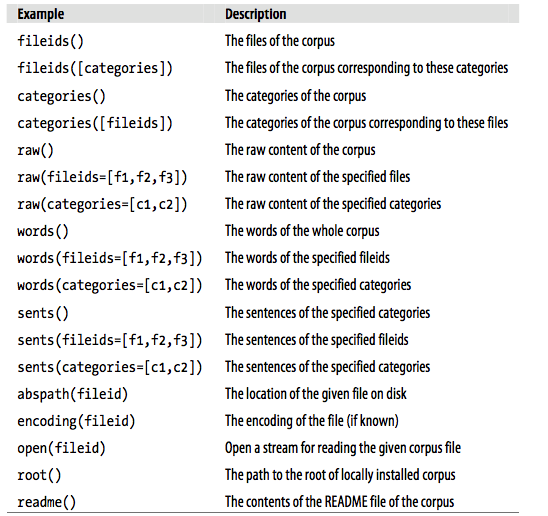
Corpus nay chứa hơn 10 ngàn tin tức và 1.3 triệu từ. Nó được chưa thành 90 chủ đề và được chưa làm 2 tập là "training" và "test".

* 1. Annotated Text Corpus:

Các Corpus được đánh dấu, NLTK cung cấp rất nhiều các Corpus đã được đánh dấu thể hiện như POP Tags, name entities, syntactic structures...

# Cấu trúc của một Corpus:

Dưới đây là một số dạng của Corpus và tương ứng với Corpus trong NLTK Một số hàm được sử dụng để làm việc với các Corpus của NLTK



1. Download an external corpus, load it, and access it:
2. Counting all the wh words in three different genres in the Brown corpus:
3. Explore frequency distribution operations on one of the web and chat text corpus files:
4. Take an ambiguous word and explore all its senses using WordNet:
5. Pick two distinct synsets and explore the concepts of hyponyms and hypernyms using WordNet:
6. Compute the average polysemy of nouns, verbs, adjectives, and adverbs according to WordNet: