**TASK 0: Understanding the problema:**

a.- what data you have

b.- What are the standard tolos and models used for that type of data.

FILES: LOCALE.blogs.txt where LOCALE is the each of the four locales en\_US, de\_DE, ru\_RU and fi\_FI. The data is from a corpus called HC Corpora:

**About The Corpora**

[www.corpora.heliohost.org](http://www.corpora.heliohost.org/)

Home Download Statistics About The Corpora Links

The corpora are collected from publicly available sources by a web crawler. The crawler checks for language, so as to mainly get texts consisting of the desired language\*.

Each entry is tagged with it's date of publication. Where user comments are included they will be tagged with the date of the main entry.

Each entry is tagged with the type of entry, based on the type of website it is collected from (e.g. newspaper or personal blog) If possible, each entry is tagged with one or more subjects based on the title or keywords of the entry (e.g. if the entry comes from the sports section of a newspaper it will be tagged with "sports" subject).

In many cases it's not feasible to tag the entries (for example, it's not really practical to tag each individual Twitter entry, though I've got some ideas which might be implemented in the future) or no subject is found by the automated process, in which case the entry is tagged with a '0'.

To save space, the subject and type is given as a numerical code.

Once the raw corpus has been collected, it is parsed further, to remove duplicate entries and split into individual lines. Approximately 50% of each entry is then deleted. Since you cannot fully recreate any entries, the entries are anonymised and this is a non-profit venture I believe that it would fall under Fair Use.

Corpus Sample

\* You may still find lines of entirely different languages in the corpus. There are 2 main reasons for that:

1. Similar languages. Some languages are very similar, and the automatic language checker could therefore erroneously accept the foreign language text.

2. "Embedded" foreign languages. While a text may be mainly in the desired language there may be parts of it in another language. Since the text is then split up into individual lines, it is possible to see entire lines written in a foreign language.

Whereas number 1 is just an out-and-out error, I think number 2 is actually desirable, as it will give a picture of when foreign language is used within the main language.

**WIKIPEDIA:**

**Natural language understanding**

Convert chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate. Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural languages concepts. Introduction and creation of language metamodel and ontology are efficient however empirical solutions. An explicit formalization of natural languages semantics without confusions with implicit assumptions such as closed world assumption (CWA) vs. open world assumption, or subjective Yes/No vs. objective True/False is expected for the construction of a basis of semantics formalization.[5]

Statistical NLP[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=4" \o "Edit section: Statistical NLP)]

*Main article: [statistical natural language processing](https://en.wikipedia.org/wiki/Stochastic_grammar" \o "Stochastic grammar)*

Statistical natural-language processing uses [stochastic](https://en.wikipedia.org/wiki/Stochastic" \o "Stochastic), [probabilistic](https://en.wikipedia.org/wiki/Probabilistic" \o "Probabilistic) and [statistical](https://en.wikipedia.org/wiki/Statistical" \o "Statistical) methods to resolve some of the difficulties discussed above, especially those which arise because longer sentences are highly ambiguous when processed with realistic grammars, yielding thousands or millions of possible analyses. Methods for disambiguation often involve the use of [corpora](https://en.wikipedia.org/wiki/Corpus_linguistics" \o "Corpus linguistics) and [Markov models](https://en.wikipedia.org/wiki/Markov_model" \o "Markov model). One among the first models of statistical natural language understanding was introduced in 1991 by [Roberto Pieraccini](https://en.wikipedia.org/wiki/Roberto_Pieraccini), Esther Levin, and Chin-Hui Lee from [Bell Laboratories](https://en.wikipedia.org/wiki/Bell_Laboratories).[[7]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-7) NLP comprises all quantitative approaches to automated [language processing](https://en.wikipedia.org/wiki/Language_processing" \o "Language processing), including probabilistic modeling, [information theory](https://en.wikipedia.org/wiki/Information_theory" \o "Information theory), and [linear algebra](https://en.wikipedia.org/wiki/Linear_algebra).[[8]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-8) The technology for statistical NLP comes mainly from [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [data mining](https://en.wikipedia.org/wiki/Data_mining), both of which are fields of [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) that involve learning from data.

ABOUT THE CORPORA