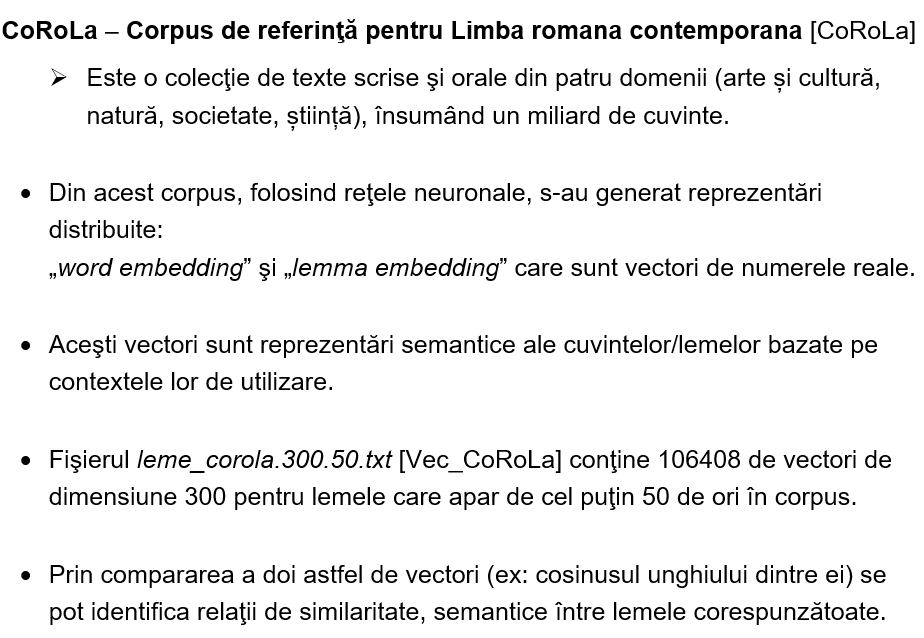
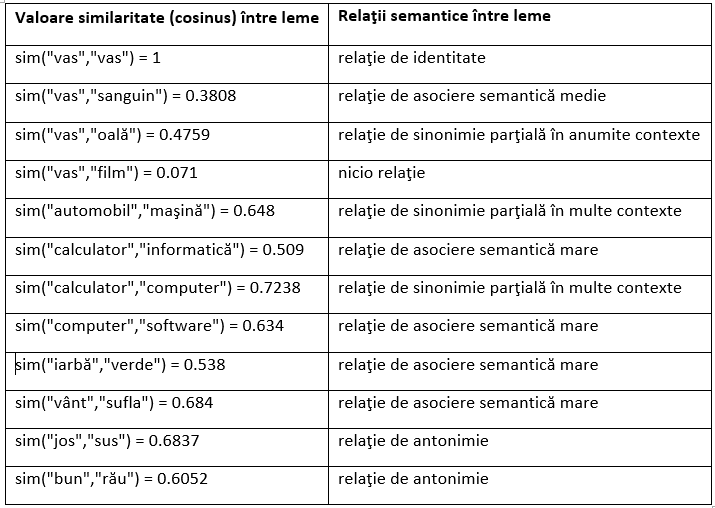
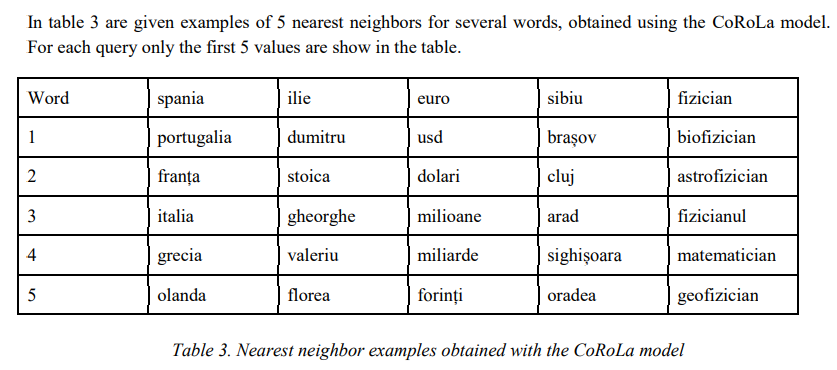
**CoRoLa –COrpus of ROmanian LAnguage-**

**Corpus de referință pentru limba română contemporană**

* It is an online resource for exploring, studying and learning about the Romanian language, both in its written and in its spoken form.
* It debuted in 2014 as a prioritary program of the Romanian Academy and was executed by Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu” of Bucharest and Institutul de Informatică Teoretică of Iași.
* The corpus contains texts from 1989 up until our days.
* The goal of this project is to offer an objective image of the Romanian language and to offer a powerful tool for researchers in the domain of linguistics, modelling and preprocessing of language, modern translation tools, intelligent indexing and recovery of textual and oral information, semantic classification of big volumes of data, information extraction, automatic text resolutions, question-and-answer systems, etc.
* The text data was cleaned of the parts which not belong into the text, such as titles, notes, bibliographies, etc. It was also checked for the correct usage of diacritics, they were added by an automatic process if not present already. Then the data was segmented on the level of sentences and with automatic and semi-automatic processes labelled metadata was added to them.
* The corpus follows the international general standards, it reflects all functional styles of a language (I.e. Imaginative, Science, Journalistic, Law, Administrative, Memoirs, Blogpost). It covers 4 general domains (Arts&Culture, Nature, Science, Society), these being subdivided into 70 categories. Here are some statistics with regards to the distribution of data in the corpus:





******

***Operations with vectors:***

***vec("rege")-vec("bărbat")+vec("femeie")=vec("regină")***

