Appendix A: Structure of the word representation space

In this project we worked with numerical representations of text dialogues, and it is therefore important to take into account the structure of these representations and how this structure can affect the analysis of these texts. We have applied different models to represent texts, but we will limit ourselves in this appendix to studying the case of GloVe, although the conclusions can be extended to other models such as BERT.

As mentioned previously, we have used the 'glove.6B' pre-trained embeddings of GloVe. This model has a vocabulary of 400,000 terms, each of which is represented by a vector with 300 components. If we analyse how these vectors are distributed in this 300-dimensional space, we can see that they are distributed anisotropically. In Figure A.1 we represent for each of the vocabulary words which is its largest component; in other words, towards which of these 300 directions each of the vectors is mainly oriented. As we can see, there are directions with a higher density of vectors than others. Figure A.2 shows the same information as a histogram, confirming the prevalence of some directions over others and therefore the anisotropy of the space of representations.

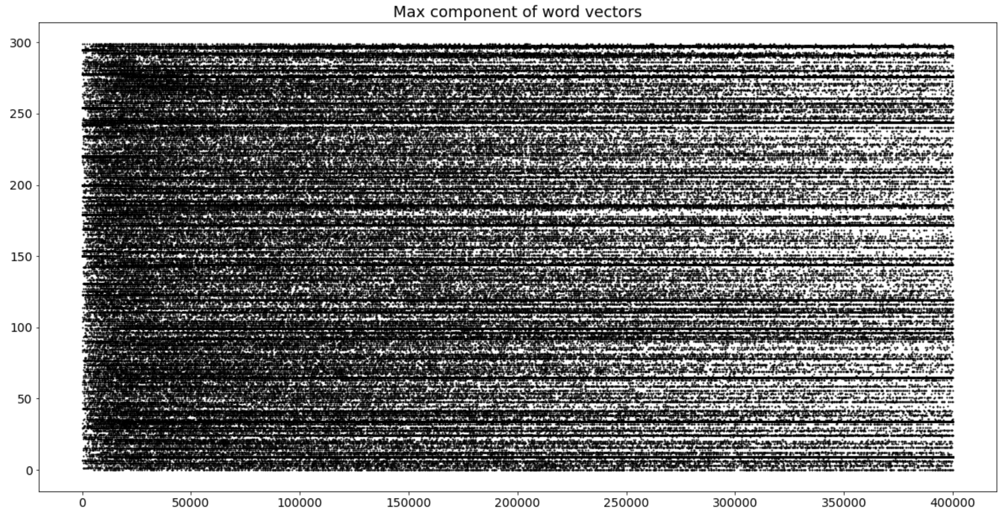

Figure A.1

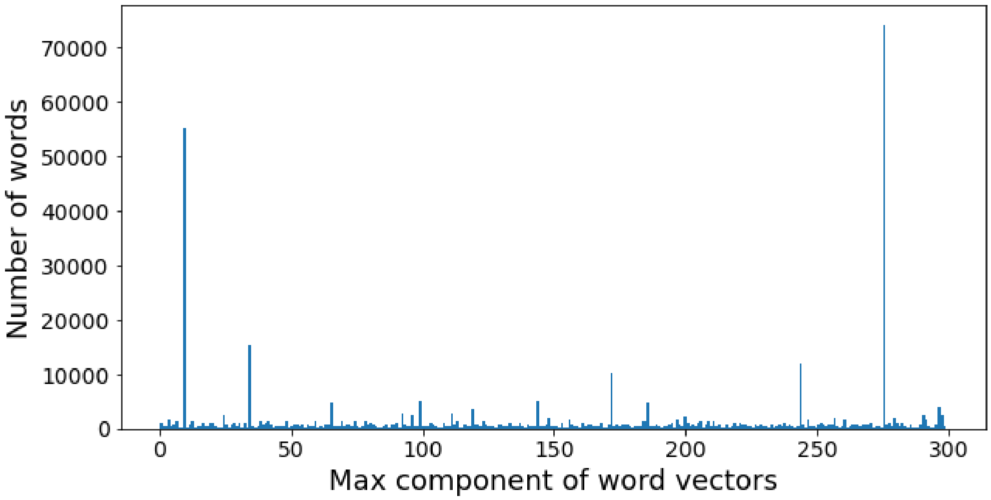

Figure A.2.

In the following, we will show the effects of such a structure when operating on long texts. To show the universality of this analysis, we will take as an example of analysis two corpora which are very different from each other and also with regard to the texts used in this project: Corpus 1 = the text of 'The Wisdom of Father Brown' by G. K. Chesterton; Corpus 2 = the text of 'Leaves of Grass' by Walt Whitman.

In Figure A.3 we plot in the top row the value of the largest and smallest components of each of the first 5000 words of Corpus 1 (we include the smallest to take into account all possible orientations of the vectors). In the bottom row we represent the same quantities, but in this case for the average vectors up to each term; for the term number n we take the average of the first n vectors. As we can see in this row, although at the beginning the extreme values oscillate, when a sufficient number of words are considered, they converge towards a value with respect to which they remain approximately stable. In Figure A.4 we can see the same behaviour in the case of Corpus 2.

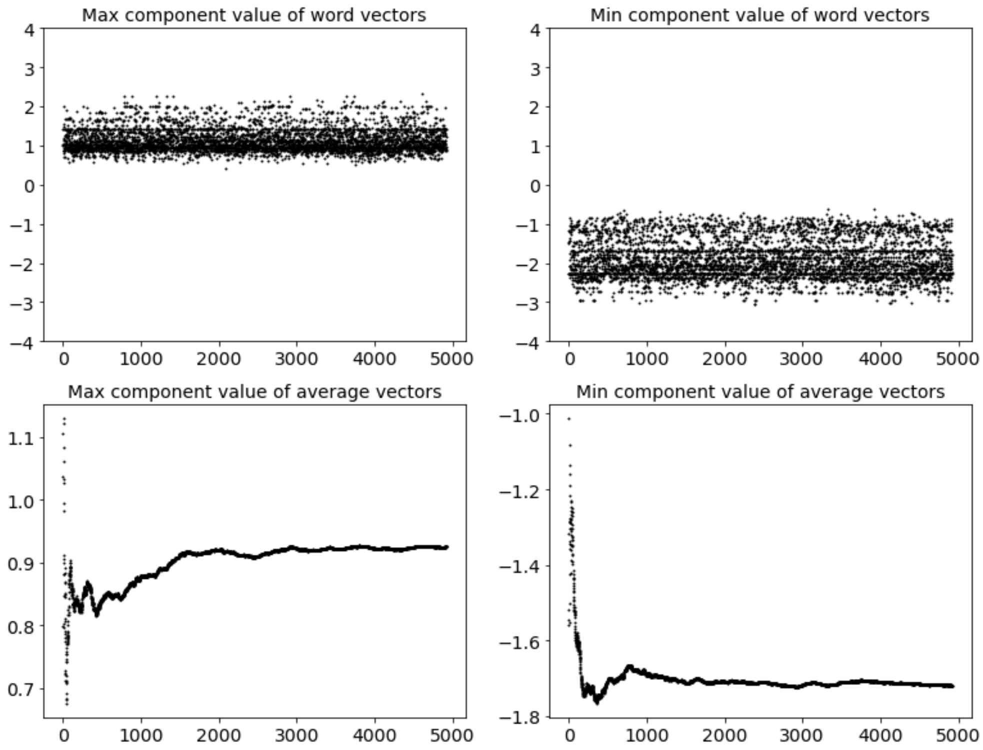

Figure A.3.

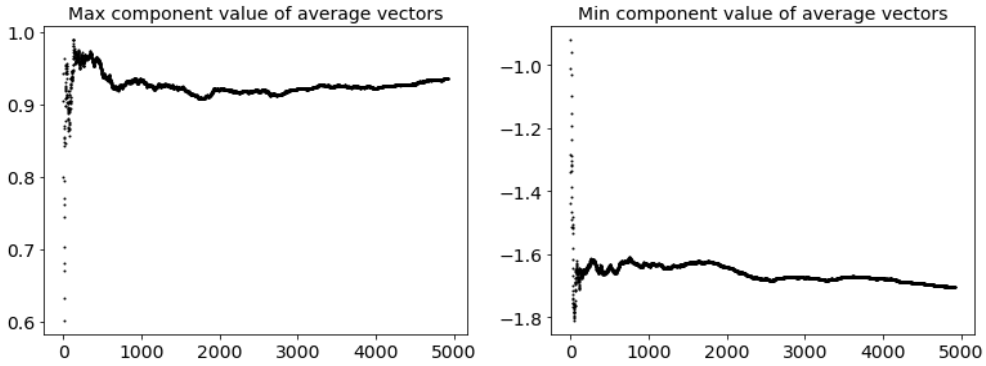

Figure A.4.

In Figure A.5 we observe again the effect of averaging over the terms, but in this case considering all the components, not only the extreme ones. The top row represents Corpus 1 and the bottom Corpus 2. Again, we observe the effect of convergence due to the anisotropy of the space of representations as we average over more vectors, and the large text similarity between the two corpora in the last column.

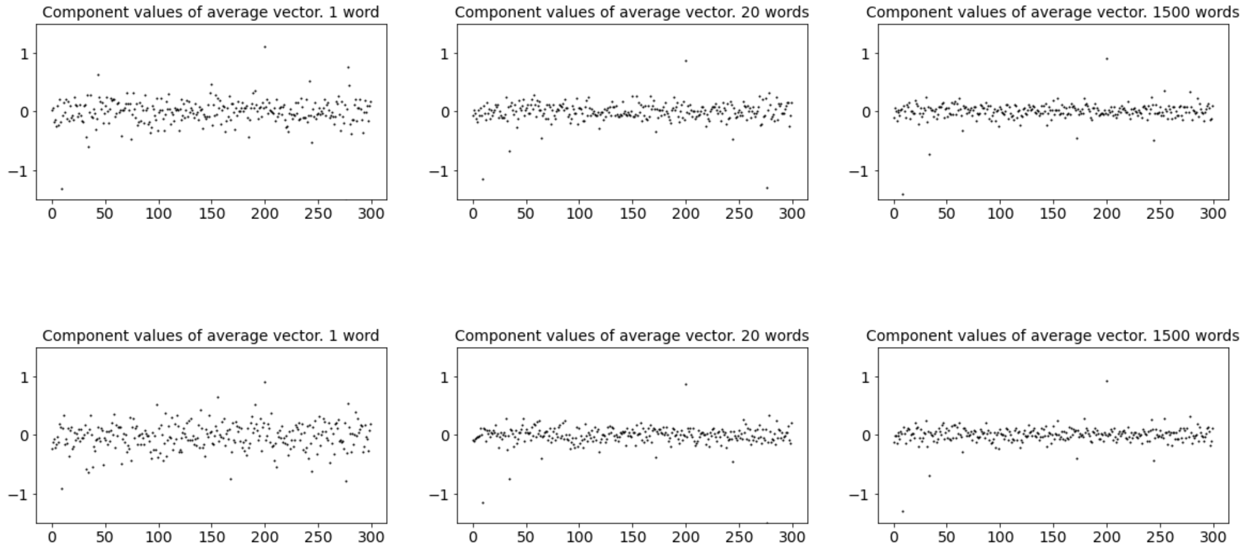

Figure A.5.

In Figure A.6 we plot the cosine similarity between the average vectors of each corpus. For the term n we compute the average vector of the first n terms in each corpus, and then we compute the cosine similarity between both resulting vectors. Although the cosine similarity measure already performs a normalisation effect, by considering only the direction of the vectors, and not their magnitude, we see here how considering sufficiently long texts has a convergence effect with respect to the values of the cosine similarity. The range of possible values will therefore reduce and converge to the unit as the amount of text considered increases. This effect must be taken into account when assessing the differences between comparisons, and especially when comparing texts of different sizes.

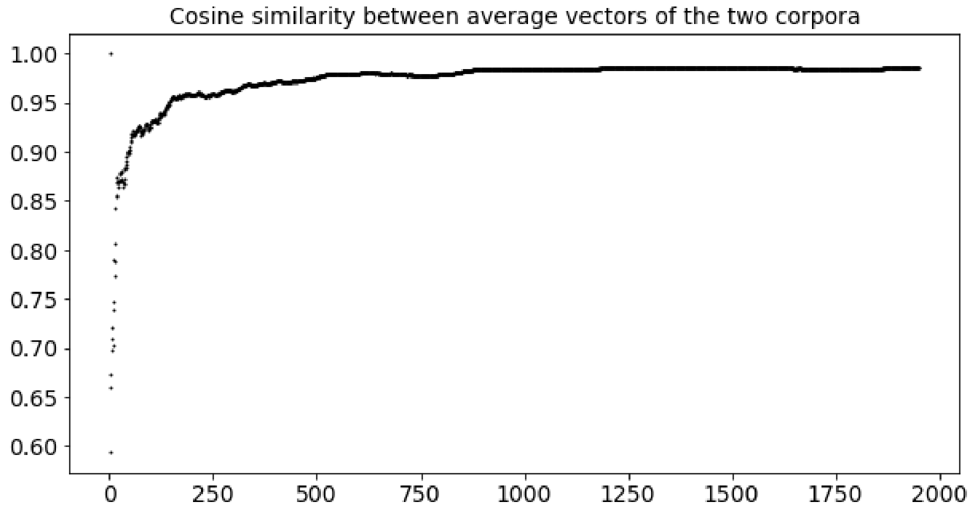

Figure A.6.