The Librarian “Amigo”: Summary

Project Design

Eight Spanish authors from the beginning of the XX century were selected for an analysis by means of supervised and unsupervised machine learning techniques. In the data set fragments of 10 sentences were the minimum considered unit. The milestones of the project were determined to be:

- **Data scrapping**.
- **Fragment generation and data storage in a single data base.**
- **Vectorization of fragments and dimensionality reduction of the vectors.**
- **Identify the author** of a given literature fragment.
    - **Binary classification**: As a first step the two authors with the largest number of fragments (~2/3 of the total data set) were studied.
    - **Multiclass classification**: all eight authors were contemplated.
- **Fragment clustering**: establish similarity relationships between fragments using clustering techniques.
- **Topic characterization of clusters.**

Data

The data was obtained from 214 public domain books from eight authors downloaded from www.gutenberg.org and www.ataun.net. Text files containing the books were scraped directly from gutenberg web site. Only a few books could be downloaded at a given time before the website blocked it for some time. The majority of the books were downloaded as pdf files from ataun.net. The files needed to be converted to text files. This was accomplished using the method pdf2txt from the pdfminer module. The books were splitted in fragments of 10 sentences after trimming each file to avoid the inclusion of indices, headers, etc. A total of 45961 fragments were obtained.

Tools and Algorithms

- **Author classification:**
    - **Vectorization:** Gaussian Naive Bayes for classification was tried on on previously vectorized fragments. Hyperparameter tuning for vectorization and dimensionality reduction algorithms was carried out taking the Naive Bayes fit as the the deciding quantity. The following methods for vectorization were used:
        - Count vectorizer with Latent Semantic Analisys (LSA)
        - Count vectorizer with Non-negative Matrix Factorization.
        - Tf-idf with LSA

        - Doc2vec.
        - Word2vec: this method was only used for the study of comparative use of words for different authors (Galdós and Baroja)
    - **Binary (Galdós/Baroja):** The tfidf/LSA method was used for training with the following methods:
        - Gaussian Naive Bayes.
        - Random Forest.
        - Support Vector Classifier.
    - **Multiclass vectorization:** tfidf/LSA method was used tor processing the fragments.
        - Gaussian Naive Bayes.
        - Random Forest.
        - Support Vector Classifier.
- **Fragment clustering:** k-means clustering with k = 2-20 was tried. A corner plot was used to for deciding the most reasonable value for k with was found to be around 7. Clustering with k = 6-8 of fragments processed using tfidf/LSA. It was checked that some clusters were consistently formed regardless of the number of components used in the vectorization and the k for the k-means method.
- **Topic characterization:** The proportion of the fragments for all authors in each cluster was quantified. The topics present for each cluster was studied by using separate tfidf/LSA vectorizations for which only fragments within a given cluster were considered.

The best test accuracies for author classification were for binary and multiclass classification were found to be 76.9% and 69.6% respectively. The result can be compared to the dummy classifiers for binary (55.3%) and multiclass (44.7%).

Review and improvements

In the present project a large fraction of the time was invested in data collection and in the study of the different vectorization methods that will give best results for Gaussian Naive Bayes classifier. As a result not so much time was left for the training of the models themselves and only a few methods with the selected vectorization method (tfidf/LSA) were made. The vectorization method was decided taking the accuracy of subsequent Gaussian Naive Bayes as indicator and it is likely that other vectorization methods would have been more adequate for other supervised methods. Also other methods such as neural networks could not be tried due to lack of time. A more detailed preprocessing of the fragments would be also necessary as a first step as it was found later in the project that some punctuation was not properly filtered out (e.g. “guerra” and “guerra.” were treated as distinct words)

As regards clustering, only k-means was tried and other potentially more adequate methods such as DBSCAN were not pursued and would be worth trying. Also further characterization of the obtained clusters particularly by sentiment analysis would be highly desirable.