Digital Text Analysis for Students of Spanish Language and Culture

Javier Sampedro
University of Florida / Loyola University of Maryland

Artifact type: Class assignments
Developed for: Upper-level undergraduate class
Audience: Spanish and Latin American studies majors
Time required: 45-minute classroom session for each activity
Method and tool: Text analysis with Voyant; location mapping and analysis with Stanford NER (Named Entity Recognition), Google Maps, and Voyant; topic modelling with Mallet and Excel.

Description:

These two class assignments were designed and tested during Fall 2018 in my course at the University of Florida on "Latin American Literature in its Context," a panoramic view of Latin American history and culture through close and distant reading of exemplary literary works from 1800-1990. (The class is recommended for upper level Spanish Major and Latin American studies majors.) Complementary to traditional literary interpretation and analysis through close reading, students participated in textual analysis projects supported by the quantitative information extracted from the works studied using computers. We enriched the critical literary debate in class with the support of simple statistical analysis that shows the structure and functioning of the Spanish language from a different perspective. During the semester the students learned to interpret frequency analysis and distribution of key terms (the class focused on place names), to use computational tools to annotate locations by listing and mapping those that appear in a text, as well as automatically generate topics.

Supporting materials:

Assignment overview for "Group Digital Annotation Project" and "Group Topic Modeling Project"

Digital Text Analysis of Latin American Short Stories

Your Name: Javier Sampedro
Your Institution: University of Florida / Loyola University of Maryland

The brief descriptions of these assignment below are only a quick reference for the student. This was a 3 hour/week course, and we spent approximately 1 hour/week learning and practicing the tools and procedures as a group.

Class Assignment A

Group Digital Annotation Project ("PALCO" in Spanish)

Objectives:

1. Creating an embeddable map containing several geographic locations [[geographic_locations]] mentioned in a literary corpus.
2. Creating a "trends" embeddable graph showing the position of these locations in the narrative timeline.

1. Automatic extraction of all occurrences of [geographic_locations] in the novel "Los de Abajo", by Mariano Azuela (1915) using Name Entity Recognizer (Stanford NLP Group).

   • Follow the instructions to **install** the Stanford-ner folders in your computer.
   • **Insert** the provided file (spanish.ancora.distsim.s512.crf.ser) into the *classifiers* folder for Spanish language processing.

   • **Run** ner-gui (graphic user interphase)

   • **Select** the correct *classifier* file from the classifiers folder for Spanish language.

   • Go to terminal and copy the entire list of tagged words.

   • Copy the list on BBEdit and delete all tagged words other than LOCATION.

   • **Copy** the entire list of words tagged as LOCATION only.

   • Using BBEdit clean the list from errors and repeated locations.

2. Visualization of [geographic_locations] using *google maps*.

   • **Sign in** to your google maps account and navigate to "your places".

   • Select "create a map"

   • Choose a **design** for your map and add at least 15 geographic locations from the NER generated list you just cleaned up using BBEdit.

• **Save** the map and make sure you can **embed** the HTML in the source code for your webpage.

3. Visualization of frequency and location of [geographic_locations] in the narrative timeline using Voyant tools.

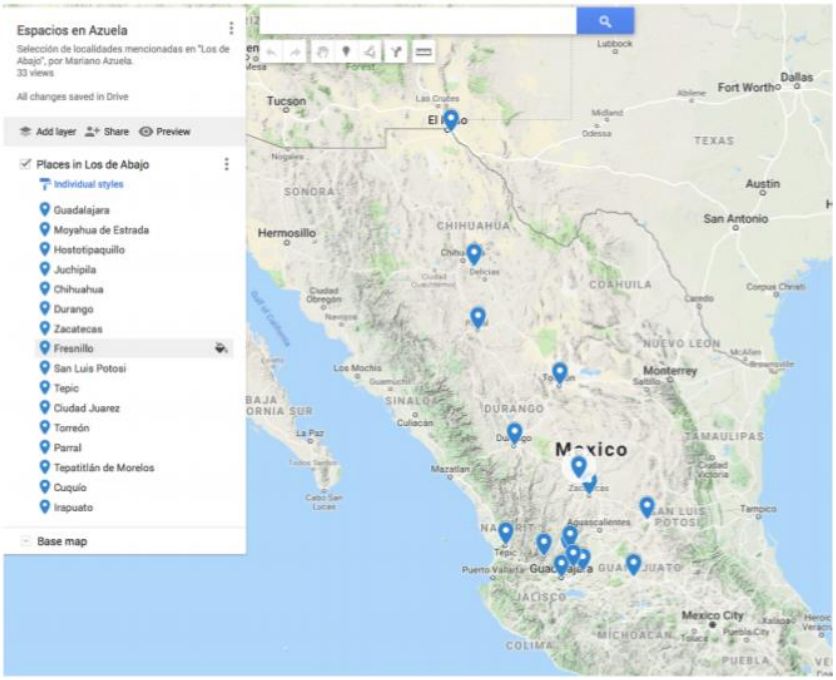
• **Open** voyant-tools.org

• **Paste** the entire corpus of the novel “Los de Abajo” onto the box. “Click Reveal”.

• Choose the graph entitled “trends” and **add all names** (one by one) tagged as LOCATION from the previous list to the *search* bar. Click reset.

• **Embed** the resulting new graph onto your webpage.

The graphs should look similar to this one on your webpage:



Class Assignment B

Group Topic Modeling Project

Objectives:

1. Creating a list of 10 topic names using the topic-words list generated by the Topic Modeling Tool.
2. Creating a graph illustrating the numerical values (percent) of each individual topic present

in every short story.

1. Processing a *clean* version of several literary short stories studied using TopicModelingTool (Mallet).

   • Download and install the TopicModelingTool graphic user interphase.

   • Set the input and out directories accordingly.

   • Set Number of topics to 10

   • In Optional Settings, upload the file containing the Spanish stopwords-list previously generated in class.

   • Upload each *cleaned* version of the story individually for processing.

   • Select “Learn Topics.” Browse the output directory for the three newly generated files.

2. Group brainstorming and writing of topic names from the automatically generated topic words by Mallet.

   • As a group, decide what title or name to give to each topic using the keywords automatically generated on each case.

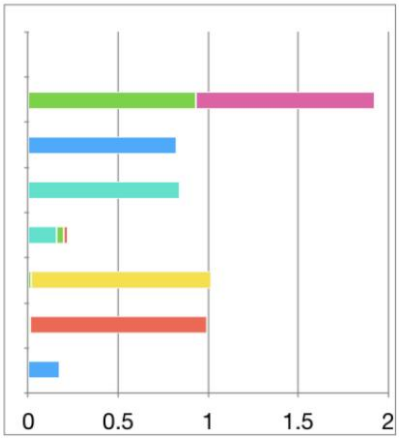
   • Change the first columns values (filename) with the title of each short story.

   • Substitute the topic wordlist for each short story with the title decided for every case.

3. Visualization of common topics across individual literary pieces using the automatically generated percentages file using Mallet and Microsoft Excel.

   • Create a graph comparing the values obtained from the TMT file in each short story. The table and graph should be embedded on your webpage.

Título	Topic o	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
El Matadero	0.00004	0.00002	0.84290	0.15696	0.00003	0.00003	0.00002
Después de las Carreras	0.92903	0.00007	0.00006	0.03889	0.01769	0.01419	0.00007
La Compuerta No. 12	0.00009	0.00005	0.00005	0.00184	0.99784	0.00008	0.00005
El Rubí	0.00013	0.00007	0.00006	0.02298	0.00010	0.97659	0.00007
Continuidad de los Parques	0.99793	0.00023	0.00022	0.00063	0.00035	0.00039	0.00024
San Antoñito	0.00005	0.82291	0.00003	0.00008	0.00004	0.00005	0.17684



Topic Id	Top Words...
0	berta manón luz vida azul mano seda espejo aliento terciopelo
1	doña pacha señoras damián damiancito fulgencita dios débora seminario aguedita
2	matadero toro había caballo juez restaurador matasiete más carne animal
3	amigo agua cutis miembros infame negro alma tirando recado
4	viejo galería mina pablo ojos minero pequeño padre rostro roca
5	mujer tierra puck diamantes gnomos oro rubí piedra
6	alma certámenes dizque reconocimiento prestigio jesucristo tales pasó manos