MSiA 490 Text Analytics
Final Project Part 1
Carson Chen

Topic: **Machine Translation (MT)** between Spanish and English

Dataset: European Parliament Proceedings Parallel Corpus 1996-2011, Spanish-English. These documents are extracted from the proceedings of the European Parliament. Matching items are labeled with corresponding document IDs and identified sentence boundaries.

Size:

- 300 MB for Spanish to English and 320 MB for English to Spanish.
- 1,965,734 sentences, 51,575,748 L1 words, and 49,093,806 English words.

Source: http://www.statmt.org/europarl/

Task: Given the nature of this dataset, I hope to develop a machine-learning system to convert formal and structured Spanish text into formal English text.