|  |  |
| --- | --- |
| **NGUYEN, Réal** | **COMP 472 SEC F** |
| **ID#275662673** | **DUE 04/12/2018** |

**MINI-PROJECT 3 REPORT**

**1. INTRODUCTION**

This report will include a description of the basic setup used to recognize sentences in English, French, and Spanish; an analysis of the results of the basic setup; a description of each experiment done on the basic setup or the corpora; and an analysis of the results of each experiment.

**2. BASIC SETUP**

*~~- Explain why you chose Spanish as OT here and the corpora used~~*

*~~- Explain what constitutes as a correctly and incorrectly tagged sentence~~*

*~~- Explain how the sentences are read for each model~~*

The basic setup constitutes a system that reads corpora as training data, to then apply the training data in a Naïve Bayes classifier to recognize which language is used in a sentence. There are three languages: French, English, and Spanish. Spanish was chosen as the third language because there are plenty of corpora freely available on the Internet, and it is in the Romance language family, like French. For the sake of brevity, French will be abbreviated as FR, English as EN, and Spanish as OT for the rest of the report, except for the conclusion.

It is important to note that for both unigrams and bigrams, punctuation and whitespace are ignored. This is especially important for bigrams. For example, the phrase “Chez Tom” would produce the bigrams ch, he, ez, zt, to, and om. The result of this is that some bigrams that appear in my models would not appear in models that account for punctuation: using the same example, the bigram zt would most likely never appear in any bigram model that accounts for punctuation and whitespace.

**2.1 Results**

*- Compare sentences w/ correct language model and actual language model side-by-side*

*- Compare unigram vs. bigram models when incorrectly tagged*

*- Use tables w/ languages and total sum probability*

*- Take sentences of interest, list them, and say why you want to talk about these*

*- Sentences incorrectly tagged in unigrams*

*- Sentences incorrectly tagged in bigrams*

*- Sentences incorrectly tagged in both*

*- Sentences using loanwords/uncommon letters*

*- Compare probability table with bigram models*

*- For each incorrect prediction, analyse and hypothesize why the model failed, and how it could be improved*

In this subsection, I will analyse the results of the language models reading the sentences using the basic setup. Because of limitations on the length of this report, this subsection will only discuss a subset of all the sentences; two for each language, one correctly classified and one incorrectly classified.

In this report, a sentence is considered incorrectly classified when at least one language model has incorrectly predicted the language a sentence is in. For example, the sentence “Birds build nests.” (EN) is incorrectly classified, because the unigram model classified as FR and the bigram model classified it as EN. Overall, 15 sentences are incorrectly classified: 5 in the default 10 sentences, and all the last 10 sentences, which are supposed to be incorrectly classified as per the handout.

From this point on, any reference to sentences in this subsection will use the form Sn where n is the sentence number.

|  |  |  |
| --- | --- | --- |
| **Sentence #** | **Sentence** | **Language** |
| S1 | Que la lumiere soit, et la lumiere fut. | FR |
| S2 | Woody Allen parle. | FR |
| S3 | The weather in Taumata[…] is lovely. | EN |
| S4 | Numismatic symmetry should not antagonize economic acme. | EN |
| S5 | Que alcahuete! | OT |
| S6 | Voy a buscar el kayak. | OT |

Each odd numbered sentence corresponds to correctly classified sentences, and each even numbered sentence corresponds to an incorrectly classified sentence.

Let us analyse the correctly classified sentences first. S1 is in proper, literary FR, so it is no surprise that it is correctly classified, as all the language models in the basic setup are trained exclusively on literature (as opposed to news articles, interview transcriptions, instruction manuals, etc.) S3, on the other hand uses a very long foreign place name but is still correctly classified. In this report, that place name is abbreviated to Taumata[…]. It is an abbreviation for Taumatawhakatangihangakoauauotamateaturipukakapikimaungahoronukupokaiwhenuakitanatahu, an 85-letter Maori language place name. The reason why it is correctly classified is that Taumata[…] contains many instances of the letters ‘w’ and ‘k’, which appear often in EN but rarely in FR or OT. The bigram (w|h) is also much more frequent in EN than in other languages because of functional words like “who”, “where”, “what”, etc. Compare these smoothed probabilities, taken from the language model output files:

|  |  |  |  |
| --- | --- | --- | --- |
| **Language** | **P(k)** | **P(w)** | **P(h|w)** |
| FR | 2.91E-04 | 0.004585144 | 0.015625 |
| EN | 0.008458685 | 0.023369316 | 0.267180174 |
| OT | 6.24E-07 | 2.14E-04 | 0.03125 |

S5 is comparatively short and is spoken language using slang. Nonetheless, S5 is accurately classified as OT in both the unigram and bigram models. For the unigram model, the approximation for OT is quite close to FR, but the bigram model gives a clear advantage to OT because the bigrams (q|u) and (u|e) are much more common in functional words, and the bigram (l|c) appears more often in lexical words. The point where (l|c) is read is the point where OT has a clear advantage over the other language models and keeps that advantage. *(insert comparison table)*

Now, let’s analyse the incorrectly classified sentences. It is obvious why S2 is classified as EN for both the unigram and bigram models: Woody Allen is a rather anglophone-sounding name, and that is reflected in the letters and bigrams used in the name. The letter ‘w’ rarely appears in FR and OT compared to EN. For the bigram model, the bigrams (w|o) and (o|o) are enough to significantly skew the prediction in EN’s favour. *(insert comparison table)*

S4 is unique among other EN sentences, in that it exclusively uses words of Greek origin, save for functional words (in this case, “should” and “not”). The unigram model incorrectly classifies this sentence as OT, but the bigram model correctly classifies it as EN mostly because of these functional words.

Finally, S6, like S2, uses uncommon letters in its words, and makes it easy for the language models to incorrectly classify it. Although all the language models use the same character set, the 26 letters of the Latin alphabet, the letter ‘k’ practically does not exist in the Spanish language, except for a few loanwords (like kayak). In fact, ‘k’ never appears in the corpus on which the OT models are trained. Compare these smoothed probabilities, from the unigram outputs: *(insert comparison table)*

Both the unigram and bigram models identify the sentence as EN. Compare these logarithmic probabilities: *(insert table)*

**3. EXPERIMENTAL SETUP**

*- Explain what each experiment is*

*- Follow same guidelines as previous section*

There are 4 experiments in this section:

**3.1 Results**

- Follow same guidelines as previous section

**4. CONCLUSION**

**5. REFERENCES**