Welcome to report #3

Hala Grefat

ID:207921008

Part 1 - tagging classes:

In this part in function *label_words()* for each word in text en_es, I classified whether it's CS or not based on the instructions, for words that are "ambiguous" we decided to classify it as not CS, I think that's because most words aren't CS, and in addition to that, the reason that it is ambiguous is ether (1) it's not in the dictionary ,which doesn't mean the word doesn't belong to that language, since the dictionary is small and doesn't include everything. (2) the word is too short or is more common in other languages (other than en / es) which isn't relevant because we're only talking about en & es words, therefore the safest bet is to clarify it as NOT CS.

Part 2 - feature vector:

In this part, using the output of previous part, and some googling I tried to figure out what features of words that help classify words into CS or NOT CS. (function: *get_feature_vector*())

#1 the example feature, words that aren't in the beginning of a sentence, but do have the first letter capitalized.

Reasons: these words are names of people/ places…, names aren't translated into deferent languages, so if someone was talking\typing in Spanish they would type the name in ENG and continue typing in Spanish, therefore a good chance names might be CS.

#2 If a word is typed in all caps.

Reason: words typed in all caps might be a shortcut name such as (SOS, KNN…) since they're names, the reason from #1 applies.

#3 words that are in both dictionaries.

Reason: I've noticed, while looking at the output result of part1, that plenty of the cases that were CS were words that were common to both languages, I noticed as well that some of these words that were classified as CS weren't really CS, for example a sentence that is in Spanish that has a word that is in both dictionaries and is detected as ENG using langdetect(),therefore a CS even though it's not really, so this feature could cause mistakes.

#4 If word contains punctuation marks, and isn't a punctuation mark.

Reason: this is for words such as i.e., A.N…, words that are short for other words, that are used universally.

#5 If word is at the beginning of a sentence.

Reason: these words are never CS so good indicator that it's not CS.

6 If word is at the end of sentence (words not ".").

Same reason as above, except it's **almost** never a CS.

#7 if word is one of (la, de, a, and, the, of, to, is, na, too, did, do, can, aka).

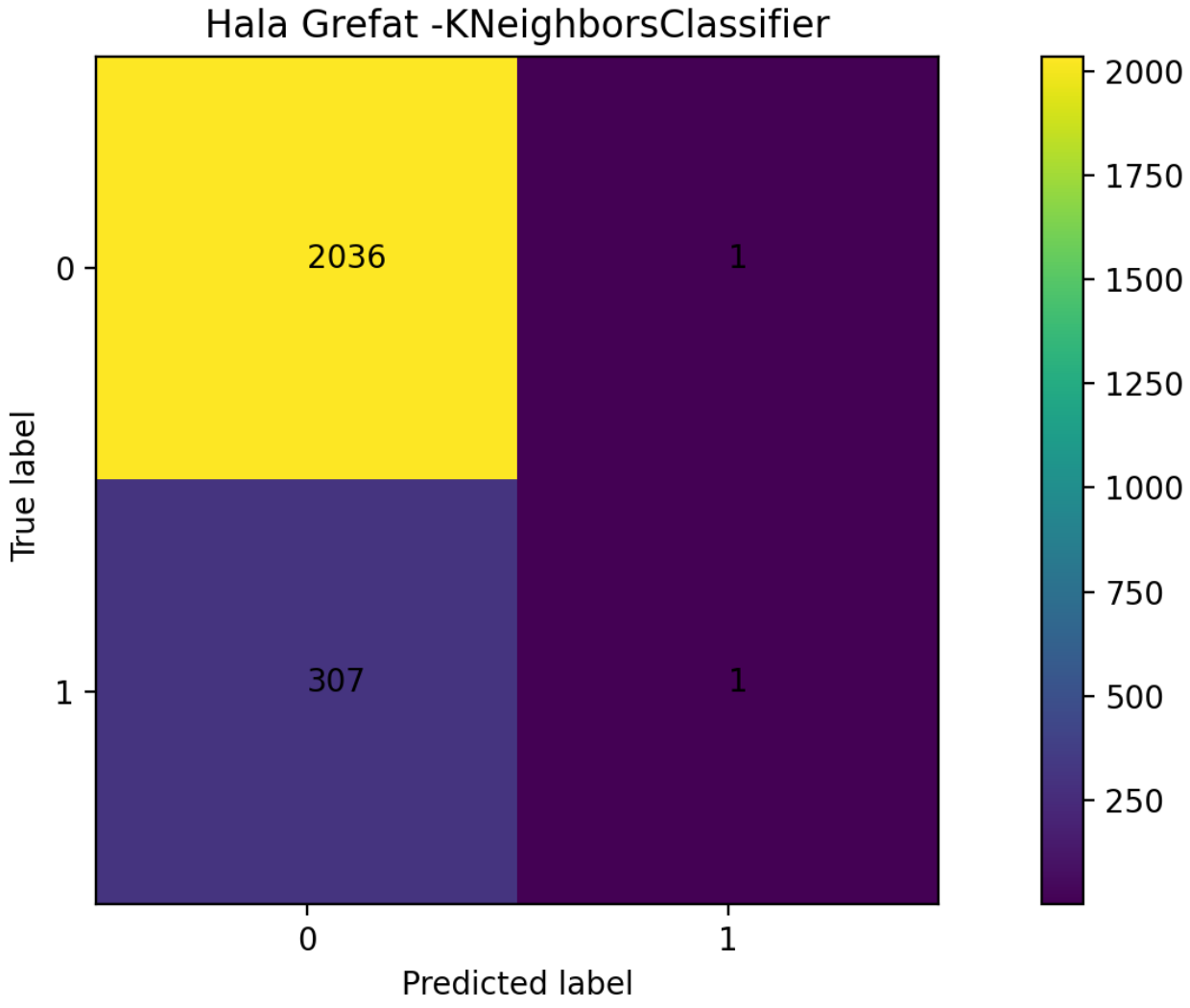
Reason: simply because I noticed they were CS many times (from parts1 output)

Part3:

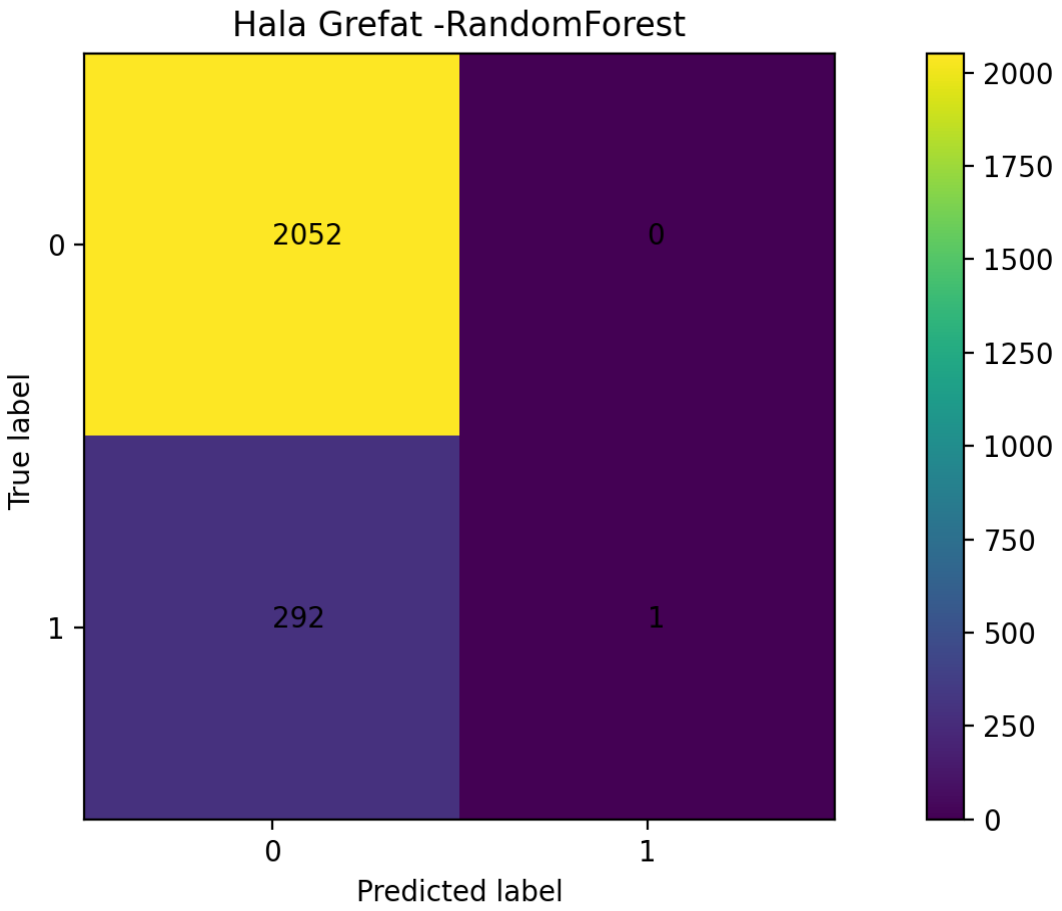
In this part, I have the data (file en_es) with each token classified, and has it's feature vector based on previous parts, I split up the data to two parts: testing and training 30:70 accordingly, using train_test_split(), with shuffle.

Using the same split, I classified the testing data, once with KNN with 10 nearest neighbors, and another with Random forest with 130 estimators.

Confusion matrix plot for KNN:



Confusion matrix plot for Random Forest:



Results discussion:

With both classifiers we can see that it is more difficult to classify words that are CS correctly, rather than words that are classified as NOT CS based on above graphs and the report. Accuracy varies from 86%-89% for both classifiers.

Answers to questions:

1. The fact that each new line is not related to previous line, made it easy to classify words that are in the beginning of line.

Since the data is from reddit there were many misspelled words, which caused to having misclassified words.

2. Features number 2,5,6. Because they don't depend on what language the data is, so they almost always work, except maybe #2 (words with first letter capitalized) doesn't work for languages that don't have capital letters but does work for the given example.
3. Having previous knowledge of the languages in data, in this case EN and ES, does help a lot, for example "aka" is a short for "also known as" and in this example it's a CS "Sinceramente no veo productos mas allá del promedio aka Best Buy Mexico (Ships from and sold by ."

Another example for how previous knowledge of the languages helps would be that names start with capital letter, which we used as a feature.

4. CS usually happens, in my opinion, when someone wants to say something but it's easier to say in other languages, for example if they're talking in their language and mention a name in another language (since names can't be translated) and continue talking in their language, this applies for names of places/TV shows/people etc.
5. "Python عفكرة عندي بشتغل من" CS because python is a name.

   "مرحبا"

   بتعرفو حدا بشغتل ب Medtronic?" CS Name

   "اااه خلص انسي...رجعت قريت المשפט في اشي كنت فاهمتو غلط"

it's just that I'm more used to using the word "משפט" more than what it is in Arabic (I'm not even sure what the word is in Arabic, yes I'm Arab).

For the first and second examples, I think my features vector would have been able to classify it correctly, the CS is a name, starts with capital letter, for the third example I don't think it would have, which is partly because it's nether EN nor ES and none of the features apply to it.

Thank you,

Have a nice day 😊!