Natural language processing

Mias Ghantous – 213461692

Faisal Omari – 325616894

Section 1)

Part 1:

we implemented a function called which takes a sentence and make the tokens as required, how did we do it?

1. If we have at the end of the token then this a regular word if the first character is a Hebrew latter and the second to last.

2. if we have Hebrew latter at the end of the token and also at the beginning then this is also a regular word

3. anything else is not a Hebrew word that we should include.

And to make the list of all the sentences we implemented which uses

Questions:

1. Min\_count: is the minimum count that would be included in the training, if it was big then we are going to learn on the most frequent words in the language, which leads to overfitting and if it were too small then we would learn on words that we don’t use them very often like names of countries which will include noise, we defined it to be in the 1 that means that every word in the corpus will be included but we would have put it bigger.

Window: is the max distance between the current word and the predict word in a sentence, small value leads to over fitting as we are not going to see a lot of repeated combinations, and not enough data on medium frequent combinations of words. And big value leads to a lot of noise as the not frequent combinations would appear in the training, we choose which we think is enough not too big and not too small.

Vector\_size: is the embedding vector size which tells how much does words are similar, too big of a number leads to overfitting because of the number of features, and too small would lead to underfitting because the feature would not be representative enough, we choose 100 which is okay.

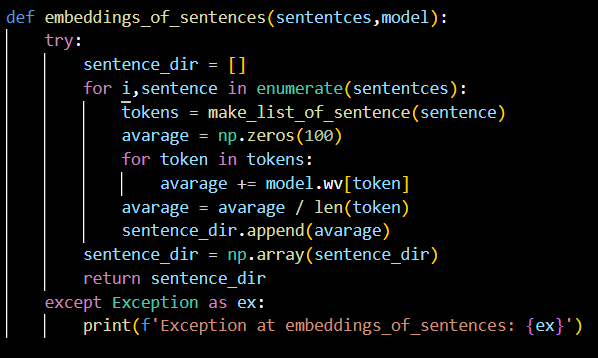
1. The problem is that we work with Hebrew and in Hebrew we have a lot of combined words that make one meaning like "בית ספר" and our corpus is not big enough to cope with this and also in Hebrew there is tokens that have many words like "וכשמהבית" which means 'and from the hose' and we created the corpus in a way that all of these words are just one token also we have words like "בצל" which can mean "in the shadow" or "onion" 2 words that have totally different meanings and in the corpus we don’t attempt to fix this.

Section 2)

Part 1. Implementation in function , we put all the words in a list and iterate throw the list.

Our results (show results after checking the format ):

Part 2. Implementation in which takes the model also the sentences and return the a dictionary from the index of the corpus to the average embedding of each sentence:



Part 3. We search the corpus for sentences that potentially we can use we choose them to be not that long has very common words in the language also sentences that can be said a lot in protocols and also can be said in a different way too, and if we are not satisfied with the result of the sentence we simply change it to another sentence.

our results (show result after cheking format):

Part 4. Our way is to first try the same word if any of the first 3 is ok then take it

Else try a similar word or a word that fits the sentence if any of the first 3 are ok then take it.

Else, try more than one word that are related.

For each word we have tried a lot and some of them got nice results and some of them we didn’t get useful results.

לחדר: השתמשנו במילה נרדפת "אולם" וכדי לקבל את ה "ל" השתמשנו ב שתבוא וקיבלנו "לאולם"

מוכנה: השתמשנו באותה מילה ובמילה שיכולה לחליף אותה כמעת תמיד "יכול" וקבלנו "יכולה".

ההסכם: השתמשנו באותה מילה וקיבלנו ההמשך שיא יכולה להחליף אבל משנה את המשמעות.

טוב: השתמשנו ב "שמש" כדי לנסות לקבל "אור" אבל הגענו למילה פחות טובה "בריא"

פותח: השתמשנו ב "מתחיל" כי היא יכולה להחליף את המילה בלי לשנות את משמעות המשפט וקיבלנו מילה משנה קצת את המשמעות "ממשיך"

שלום: זו המילה הקשה ביותר נסינו לקבל "אהלן" אבל לא הגענו לשום דבר, ואז סמנו מלים שמשתמשים בהם עם חבריים ("תודה", "חברי","רבותי") עם "שלום" וקיבלנו עמיתי, שהיא יכולה להופיע במשפט הזה.

היקר: השתמשנו ב "הטוב" כי היא מילה שיכולה להחליף וקיבלנו "גדול", היא משנה קצת את המשמעות אבל זה מה שיכלנו לעשות.

בשנה: השתמשנו באותה מילה וקיבלנו "השנה".

We can see that in all of the words we have grammatical errors, some of the words changed the meaning of the words a bit.

Our results: (show results with a photo after checking the format)

Questions:

1. we assumed that related words would be more similar like:

כנסת, ממשלה: קבלה 0.45 שהוא ערך גבוה וזה הגיוני כי כנסת וממשלה מאוד כשורים.

ישראל, כנסת: קיבלה 0.11 הקשר של הכנסת הוא יותר גדול.

ישראל, חבר: אין כשר ולכן נמוך 0.08

כל מילה עם עצמה: זה הכי טבעי שכל אחת תקבל 1.0 כי המילה היא הכי קרובה לעצמה

ישראל שולחן: אין דמיון ולכן ערך נמוך 0.09

but there are also words that have almost no relation, but got high value like and we didn’t expect that, also words that are expected them to be more related because we are talking about Knesset corpus, like:

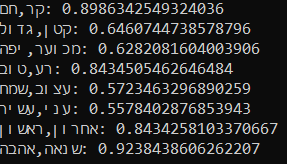
חבר, כנסת: יש הרבה הופעות של "חבר כנסת" אבל קבלנו ערך 0.2 שהוא אמור להיות יותר, זה יכול להיות בגלל ש כנסת הופיעה יותר מן "חבר כנסת" .

שולחן, חבר: קיבלנו ערך 0.6 שזה מאוד גדול והם אינם דומים בכלל.

שלום, שולחן: פחות קיבלנו 0.31 צפינו שיהיה.

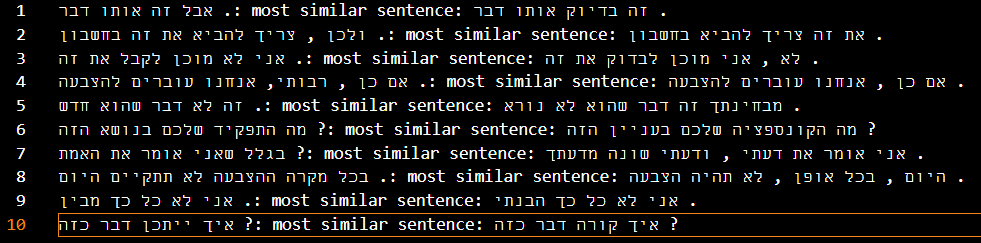
Why this happen? We don’t have big enough corpus, if we had one we would have got better results.

1. We would have small destine because a lot time, if we have 2 words (antonyms) and have a sentence that use one of the words (let's say the first one), then we can replace the first word with the second word and have a valid sentence.
2. We tried a lot of words and here are the results:



As we can see that all of the antonyms has a score over 0.55 which means that they are very similar, which is as we explained in the previous question, and even got words over 0.84.

Here are our results again:



We see that, the sentences 1,2,4,8,9,10 have the same meaning.

In sentence 3: the both are talking about ideas that needs to be checked (the second sentence) and to be agreed up on (the first sentence) and both rejection.

Sentence 5: both of the sides are describing a subject ("דבר").

Sentence 6: the word "עניין" is similar to "נושא" also in the both of the sentences the talker is talking in the same ("("גוף תחבירי and it's ("("אתם, we can see that by using ("שלכם")

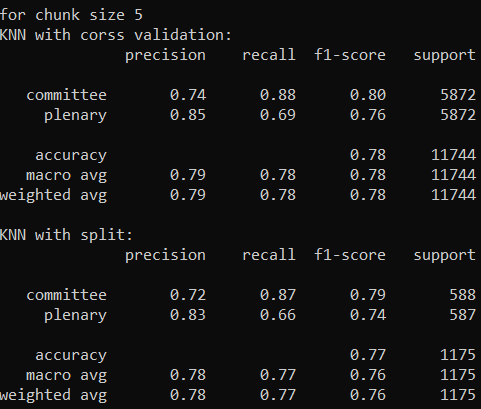
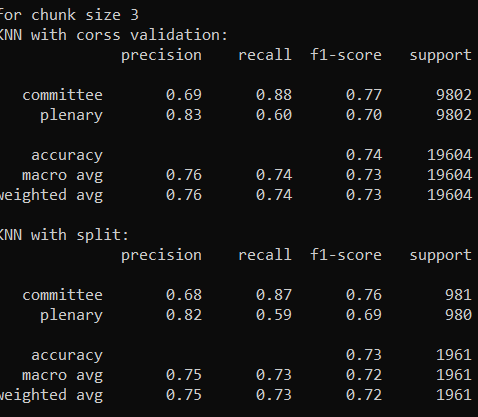
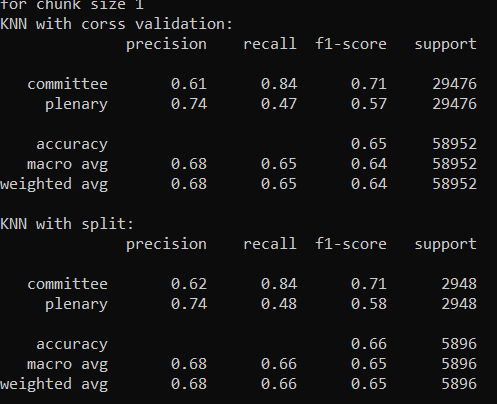
Sentence 7: both of the sentence give the something on the talker point of view.

As we can see we have very nice results and that is because of the way we choose the sentences, for each sentence, the words of the sentence have a high probability to appear together and so they have high probability to appear another time in the corpus, and that’s why we can find similar sentences.

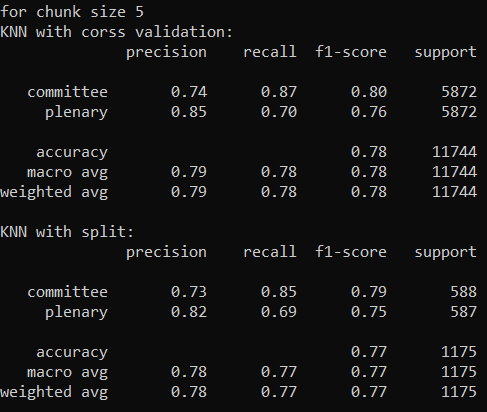
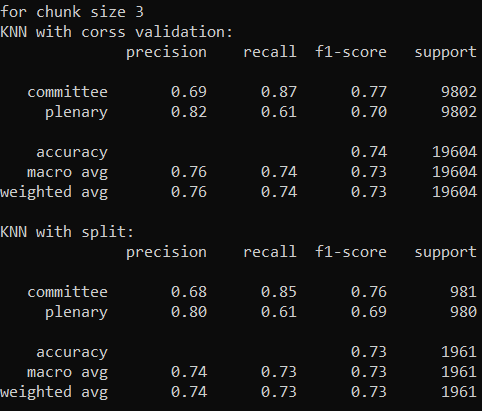
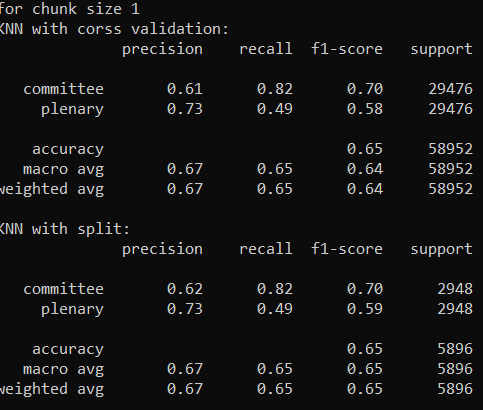
Section 3)

We tired 25, 100,50 and 10 and the got similar accuracy so we choose to be in the middle and here are the results because we don’t want too big or too small of a k:

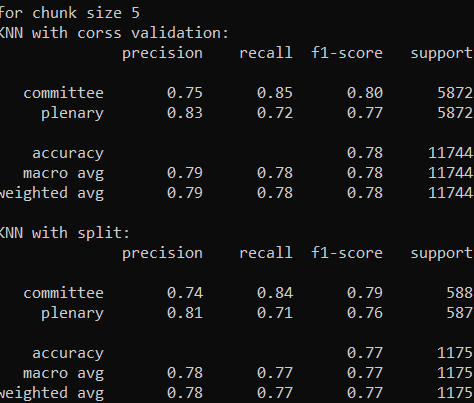
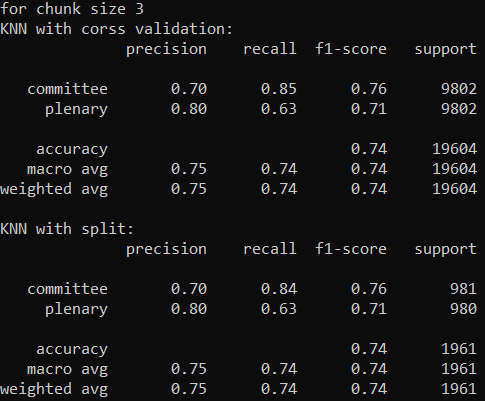
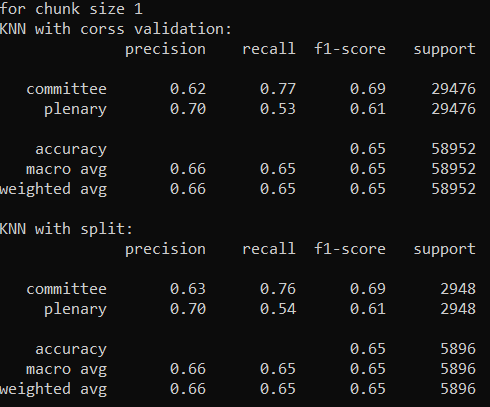
For 100:



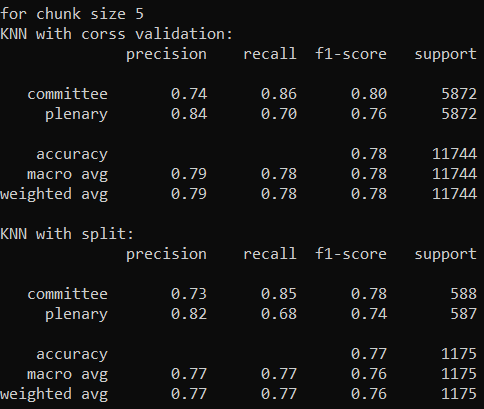
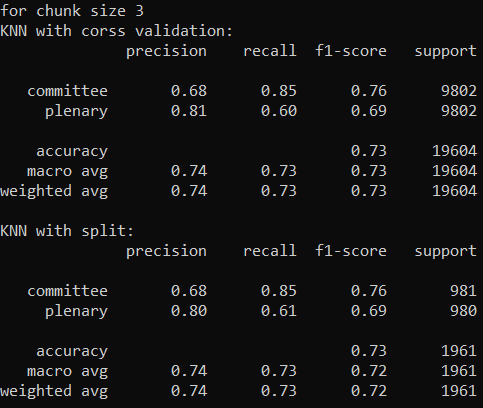
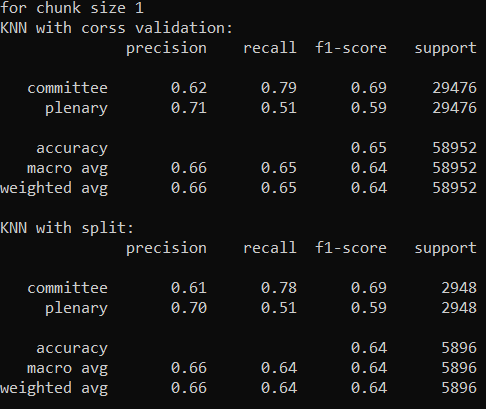
For 50:



For 25:

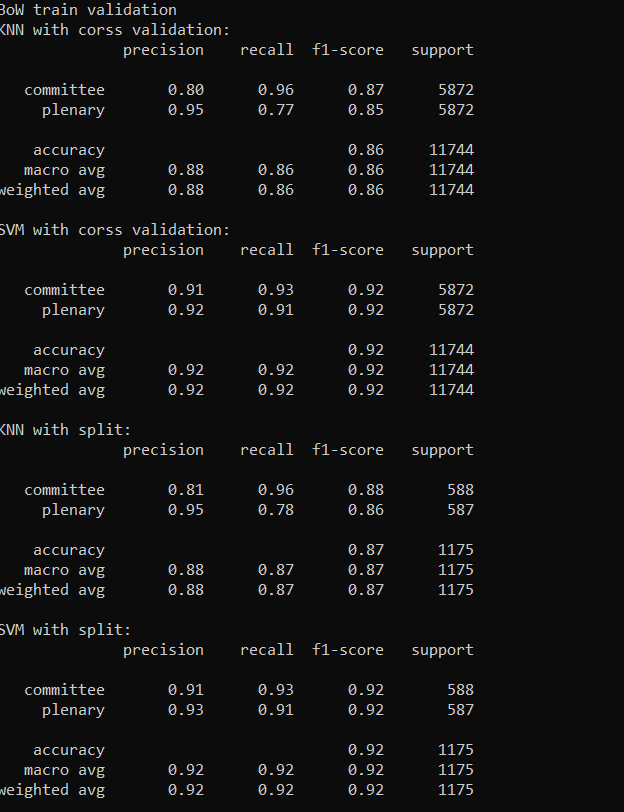
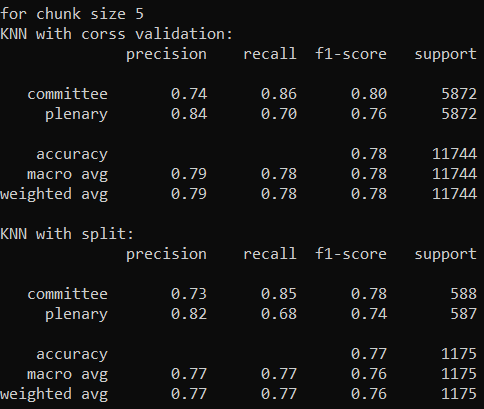


For 10:



Questions:

1. We got worst results here are our results from the past home work:

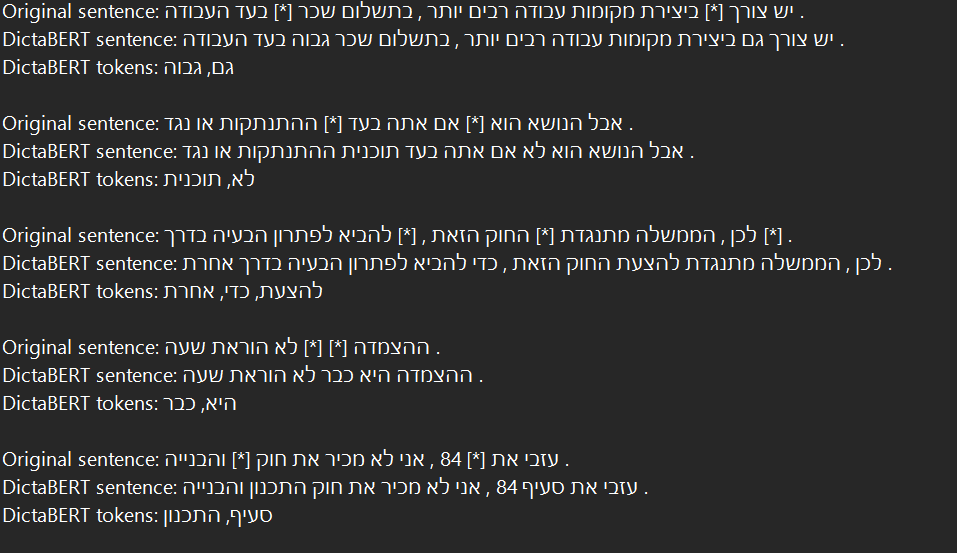
1. It could be because that the current feature vector is not as representative of the class as the previous feature vector, and that is because previously we had feature for each word, but now 100 features are going to represent all the words.
2. We always get that size 1 is worse than 3, and 3 worse than 5, why?

if we have small chunk size then we would have bigger feature matrix which is worst for run time, also each chunk would have features that doesn’t represent the types good enough, because the "feature power" would split among the smaller chunks rather than be powerful in one chunk, which make the classification task harder.

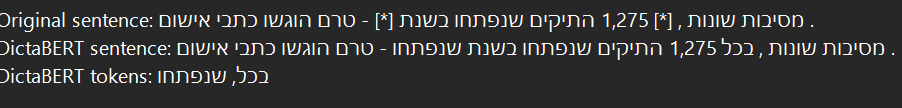
Section 4)

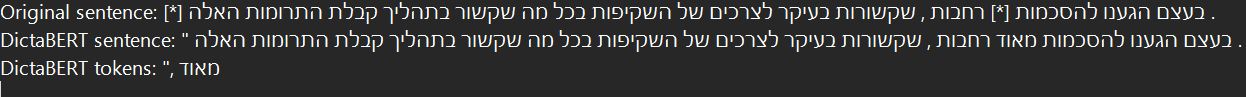
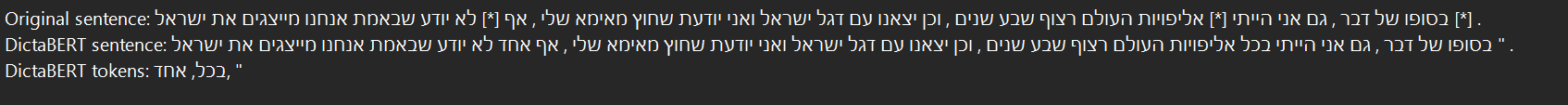
1. Form grammar perspective we see that most the sentences have good grammar

And most of them also make since like:



but here an example on a sentence that didn’t make since:



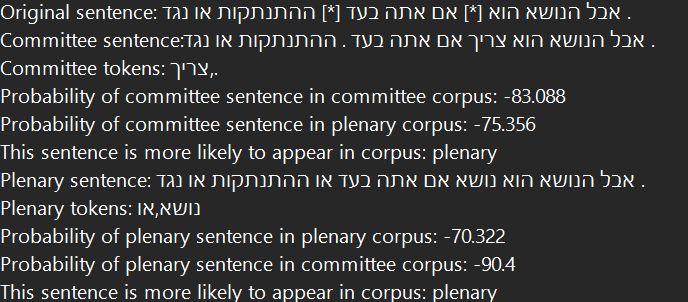


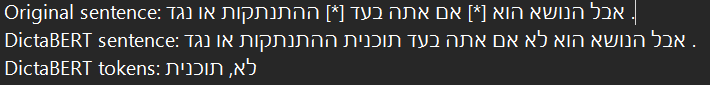
The first sentence the token "שנפתחו" is a token a good token from grammar stand point also has no meaning.

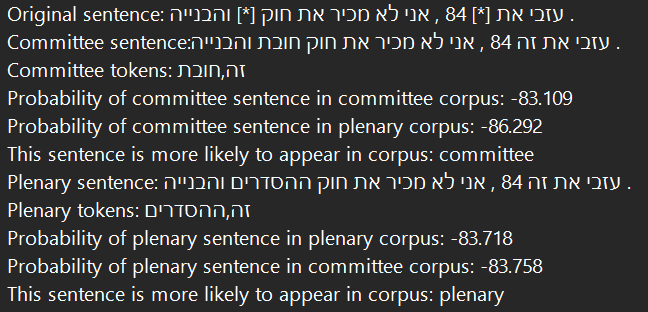
The second and third,

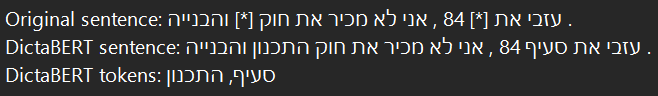
we can see that the model predicted ' \" ' without there to be any closing one which is not what grammar say.

1. Most of the results in the previous Homework were punctuation mark but now their actual words, and even when the previous model got words our current model gives more accurate predictions like:





Other example:



As we see the current model is better than the previous one.

1. As we can see above the model didn’t work good on some sentences, reasons may be this happens:

We see that in the first sentence the token must be a number (year number) it could be that he does not predict a numbers well because there are endless possibilities.

As for the second and third, if he can see all the sentence then he would know there is no another ' \" ' which make him predict another token.