Natural language processing

Mias Ghantous – 213461692

Faisal Omari – 325616894

Section 1)

Part 1:

We have implemented a function called $make_list_of_sentence$
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
$make_list$ which uses $make_list_of_sentence$

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 $Window = 5$ 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.

2. 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 $Section2_part1$, we put all the words in a list and iterate throw the list.

Our results:

Knesset_similar_words.txt:

knesset_dictabert.py | knesset_similar_words.txt

```
(0.09953677654266357,שולחן),(0.37252214550971985,שלום),(0.08438344299793243,חבר),(0.3935275673866272,ממשלה),(0.11721046268939972,כנסת),(0.9999999403953552,ישראל) :ישראל
(0.23203295469284058,שולחן),(0.4252460300922394,שלום),(0.22623348236083984,חבר),(0.4523324966430664,ממשלה),(0.9999999403953552,כנסת),(0.11721046268939972,ישראל) :כנסת
(0.23895156383514404,שולחן),(0.30447089672088623,שלום),(0.18306148052215576,חבר),(1.0,ממשלה),(0.4523324966430664,כנסת),(0.3935275673866272,ישראל) :ממשלה
(0.6516565084457397,שולחן),(0.5479722619056702,שלום),(1.0,חבר),(0.18306148052215576,ממשלה),(0.22623348236083984,כנסת),(0.08438344299793243,ישראל) :חבר
(0.28298771381378174,שולחן),(1.0,שלום),(0.5479722619056702,חבר),(0.30447089672088623,ממשלה),(0.4252460300922394,כנסת),(0.37252214550971985,ישראל) :שלום
(1.0,שולחן),(0.28298771381378174,שלום),(0.6516565084457397,חבר),(0.23895156383514404,ממשלה),(0.23203295469284058,כנסת),(0.09953677654266357,ישראל) :שולחן
```

Part 2. Implementation in $embeddings_of_sentences$ 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:

```python
def embeddings_of_sentences(sententces,model):
    try:
        sentence_dir = []
        for i,sentence in enumerate(sententces):
            tokens = make_list_of_sentence(sentence)
            avarage = np.zeros(100)
            for token in tokens:
                avarage += model.wv[token]
            avarage = avarage / len(token)
            sentence_dir.append(avarage)
        sentence_dir = np.array(sentence_dir)
        return sentence_dir
    except Exception as ex:
        print(f'Exception at embeddings_of_sentences: {ex}')
```

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:

Knesset_similar_sentences.txt:

knesset_dictabert.py | knesset_similar_sentences.txt

```
אבל זה אותו דבר .: most similar sentence: זה בדיוק אותו דבר .
ולכן , צריך להביא את זה בחשבון .: most similar sentence: את זה צריך להביא בחשבון .
אני לא מוכן לקבל את זה .: most similar sentence: אני מוכן לבדוק את זה , לא .
אם כן , רבותי, אנחנו עוברים להצבעה .: most similar sentence: אנחנו עוברים להצבעה , אם כן .
זה לא דבר שהוא חדש .: most similar sentence: מבחינתך זה דבר שהוא לא נורא .
מה התפקיד שלכם בנושא הזה ?: most similar sentence: מה הקונספציה שלכם בעניין הזה ?
בגלל שאני אומר את האמת ?: most similar sentence: ודעתי שונה מדעתך , אני אומר את דעתי .
בכל מקרה ההצבעה לא תתקיים היום .: most similar sentence: לא תהיה הצבעה , בכל אופן , היום .
אני לא כל כך מבין .: most similar sentence: אני לא כל כך הבנתי .
איך ייתכן דבר כזה ?: most similar sentence: איך קורה דבר כזה ?
```

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:

red_words_sentences.txt:

knesset_dictabert.py red_words_sentences.txt

```
1  ברוכים הבאים , הכנסו בבקשה לחדר .: ברוכים הבאים , הכנסו בבקשה לאולם .
2  אני מוכנה להאריך את ההסכם באותם תנאים .: אני יכולה להאריך את המהלך באותם תנאים .
3  בוקר טוב , אני פותח את הישיבה .: בוקר בריא , אני ממשיך את הישיבה .
4  שלום , הערב התבשרנו שחברינו היקר לא ימשיך איתנו בשנה הבאה .: עמיתי , הערב התבשרנו שחברינו הגדול לא ימשיך איתנו השנה הבאה .
```

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.

2. 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.

3. We tried a lot of words and here are the results:

   ```
   חם,קר: 0.8986342549324036
   גדול,קטן: 0.6460744738578796
   יפה,מכוער: 0.6282081604003906
   טוב,רע: 0.8434505462646484
   שמח,עצוב: 0.5723463296890259
   יש,עני: 0.5578402876853943
   ראשון,אחרון: 0.8434258103370667
   אהבה,שנאה: 0.9238438606262207
   ```

   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.

4. Here are our results again:

```
1   אבל זה אותו דבר .: most similar sentence: זה בדיוק אותו דבר .
2   צריך להביא את זה בחשבון , ולכן .: most similar sentence: את זה צריך להביא בחשבון .
3   אני לא מוכן לקבל את זה .: most similar sentence: אני מוכן לבדוק את זה , לא .
4   רבותי, אנחנו עוברים להצבעה , אם כן .: most similar sentence: אנחנו עוברים להצבעה , אם כן .
5   זה לא דבר שהוא חדש .: most similar sentence: מבחינתך זה דבר שהוא לא נורא .
6   מה התפקיד שלכם בנושא הזה ?: most similar sentence: מה הקונספציה שלכם בעניין הזה ?
7   בגלל שאני אומר את האמת ?: most similar sentence: אני אומר את דעתי , ודעתי שונה מדעתך .
8   בכל מקרה ההצבעה לא תתקיים היום .: most similar sentence: לא תהיה הצבעה , בכל אופן , היום .
9   אני לא כל כך מבין .: most similar sentence: אני לא כל כך הבנתי .
10  איך ייתכן דבר כזה ?: most similar sentence: איך קורה דבר כזה ?
```

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 $k = 50$ and here are the results because we don't want too big or too small of a k:

For 100:

```
for chunk size 1
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.61      0.84      0.71     29476
     plenary       0.74      0.47      0.57     29476

    accuracy                           0.65     58952
   macro avg       0.68      0.65      0.64     58952
weighted avg       0.68      0.65      0.64     58952

KNN with split:
              precision    recall  f1-score   support

   committee       0.62      0.84      0.71      2948
     plenary       0.74      0.48      0.58      2948

    accuracy                           0.66      5896
   macro avg       0.68      0.66      0.65      5896
weighted avg       0.68      0.66      0.65      5896
```

```
for chunk size 3
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.69      0.88      0.77      9802
     plenary       0.83      0.60      0.70      9802

    accuracy                           0.74     19604
   macro avg       0.76      0.74      0.73     19604
weighted avg       0.76      0.74      0.73     19604

KNN with split:
              precision    recall  f1-score   support

   committee       0.68      0.87      0.76       981
     plenary       0.82      0.59      0.69       980

    accuracy                           0.73      1961
   macro avg       0.75      0.73      0.72      1961
weighted avg       0.75      0.73      0.72      1961
```

```
for chunk size 5
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.74      0.88      0.80      5872
     plenary       0.85      0.69      0.76      5872

    accuracy                           0.78     11744
   macro avg       0.79      0.78      0.78     11744
weighted avg       0.79      0.78      0.78     11744

KNN with split:
              precision    recall  f1-score   support

   committee       0.72      0.87      0.79       588
     plenary       0.83      0.66      0.74       587

    accuracy                           0.77      1175
   macro avg       0.78      0.77      0.76      1175
weighted avg       0.78      0.77      0.76      1175
```

For 50:

```
for chunk size 1
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.61      0.82      0.70     29476
     plenary       0.73      0.49      0.58     29476

    accuracy                           0.65     58952
   macro avg       0.67      0.65      0.64     58952
weighted avg       0.67      0.65      0.64     58952

KNN with split:
              precision    recall  f1-score   support

   committee       0.62      0.82      0.70      2948
     plenary       0.73      0.49      0.59      2948

    accuracy                           0.65      5896
   macro avg       0.67      0.65      0.65      5896
weighted avg       0.67      0.65      0.65      5896
```

```
for chunk size 3
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.69      0.87      0.77      9802
     plenary       0.82      0.61      0.70      9802

    accuracy                           0.74     19604
   macro avg       0.76      0.74      0.73     19604
weighted avg       0.76      0.74      0.73     19604

KNN with split:
              precision    recall  f1-score   support

   committee       0.68      0.85      0.76       981
     plenary       0.80      0.61      0.69       980

    accuracy                           0.73      1961
   macro avg       0.74      0.73      0.73      1961
weighted avg       0.74      0.73      0.73      1961
```

```
for chunk size 5
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.74      0.87      0.80      5872
     plenary       0.85      0.70      0.76      5872

    accuracy                           0.78     11744
   macro avg       0.79      0.78      0.78     11744
weighted avg       0.79      0.78      0.78     11744

KNN with split:
              precision    recall  f1-score   support

   committee       0.73      0.85      0.79       588
     plenary       0.82      0.69      0.75       587

    accuracy                           0.77      1175
   macro avg       0.78      0.77      0.77      1175
weighted avg       0.78      0.77      0.77      1175
```

For 25:

```
for chunk size 1
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.62      0.77      0.69     29476
     plenary       0.70      0.53      0.61     29476

    accuracy                           0.65     58952
   macro avg       0.66      0.65      0.65     58952
weighted avg       0.66      0.65      0.65     58952

KNN with split:
              precision    recall  f1-score   support

   committee       0.63      0.76      0.69      2948
     plenary       0.70      0.54      0.61      2948

    accuracy                           0.65      5896
   macro avg       0.66      0.65      0.65      5896
weighted avg       0.66      0.65      0.65      5896
```

```
for chunk size 3
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.70      0.85      0.76      9802
     plenary       0.80      0.63      0.71      9802

    accuracy                           0.74     19604
   macro avg       0.75      0.74      0.74     19604
weighted avg       0.75      0.74      0.74     19604

KNN with split:
              precision    recall  f1-score   support

   committee       0.70      0.84      0.76       981
     plenary       0.80      0.63      0.71       980

    accuracy                           0.74      1961
   macro avg       0.75      0.74      0.74      1961
weighted avg       0.75      0.74      0.74      1961
```

```
for chunk size 5
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.75      0.85      0.80      5872
     plenary       0.83      0.72      0.77      5872

    accuracy                           0.78     11744
   macro avg       0.79      0.78      0.78     11744
weighted avg       0.79      0.78      0.78     11744

KNN with split:
              precision    recall  f1-score   support

   committee       0.74      0.84      0.79       588
     plenary       0.81      0.71      0.76       587

    accuracy                           0.77      1175
   macro avg       0.78      0.77      0.77      1175
weighted avg       0.78      0.77      0.77      1175
```

For 10:

```
for chunk size 1
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.62      0.79      0.69     29476
     plenary       0.71      0.51      0.59     29476

    accuracy                           0.65     58952
   macro avg       0.66      0.65      0.64     58952
weighted avg       0.66      0.65      0.64     58952

KNN with split:
              precision    recall  f1-score   support

   committee       0.61      0.78      0.69      2948
     plenary       0.70      0.51      0.59      2948

    accuracy                           0.64      5896
   macro avg       0.66      0.64      0.64      5896
weighted avg       0.66      0.64      0.64      5896
```

```
for chunk size 3
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.68      0.85      0.76      9802
     plenary       0.81      0.60      0.69      9802

    accuracy                           0.73     19604
   macro avg       0.74      0.73      0.73     19604
weighted avg       0.74      0.73      0.73     19604

KNN with split:
              precision    recall  f1-score   support

   committee       0.68      0.85      0.76       981
     plenary       0.80      0.61      0.69       980

    accuracy                           0.73      1961
   macro avg       0.74      0.73      0.72      1961
weighted avg       0.74      0.73      0.72      1961
```

```
for chunk size 5
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.74      0.86      0.80      5872
     plenary       0.84      0.70      0.76      5872

    accuracy                           0.78     11744
   macro avg       0.79      0.78      0.78     11744
weighted avg       0.79      0.78      0.78     11744

KNN with split:
              precision    recall  f1-score   support

   committee       0.73      0.85      0.78       588
     plenary       0.82      0.68      0.74       587

    accuracy                           0.77      1175
   macro avg       0.77      0.77      0.76      1175
weighted avg       0.77      0.77      0.76      1175
```

Questions:

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

```
BoW train validation
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.80      0.96      0.87      5872
     plenary       0.95      0.77      0.85      5872

    accuracy                           0.86     11744
   macro avg       0.88      0.86      0.86     11744
weighted avg       0.88      0.86      0.86     11744

SVM with corss validation:
              precision    recall  f1-score   support

   committee       0.91      0.93      0.92      5872
     plenary       0.92      0.91      0.92      5872

    accuracy                           0.92     11744
   macro avg       0.92      0.92      0.92     11744
weighted avg       0.92      0.92      0.92     11744

KNN with split:
              precision    recall  f1-score   support

   committee       0.81      0.96      0.88       588
     plenary       0.95      0.78      0.86       587

    accuracy                           0.87      1175
   macro avg       0.88      0.87      0.87      1175
weighted avg       0.88      0.87      0.87      1175

SVM with split:
              precision    recall  f1-score   support

   committee       0.91      0.93      0.92       588
     plenary       0.93      0.91      0.92       587

    accuracy                           0.92      1175
   macro avg       0.92      0.92      0.92      1175
weighted avg       0.92      0.92      0.92      1175
```

```
for chunk size 5
KNN with corss validation:
              precision    recall  f1-score   support

   committee       0.74      0.86      0.80      5872
     plenary       0.84      0.70      0.76      5872

    accuracy                           0.78     11744
   macro avg       0.79      0.78      0.78     11744
weighted avg       0.79      0.78      0.78     11744

KNN with split:
              precision    recall  f1-score   support

   committee       0.73      0.85      0.78       588
     plenary       0.82      0.68      0.74       587

    accuracy                           0.77      1175
   macro avg       0.77      0.77      0.76      1175
weighted avg       0.77      0.77      0.76      1175
```

2. 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.

3. 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:

Original sentence: יש צורך [*] ביצירת מקומות עבודה רבים יותר , בתשלום שכר [*] בעד העבודה .
DictaBERT sentence: יש צורך גם ביצירת מקומות עבודה רבים יותר , בתשלום שכר גבוה בעד העבודה .
DictaBERT tokens: גם,גבוה

Original sentence: אבל הנושא הוא [*] אם אתה בעד [*] ההתנתקות או נגד .
DictaBERT sentence: אבל הנושא הוא לא אם אתה בעד תוכנית ההתנתקות או נגד .
DictaBERT tokens: לא, תוכנית

Original sentence: לכן , הממשלה מתנגדת [*] להצעת החוק הזאת , כדי להביא לפתרון הבעיה בדרך [*] .
DictaBERT sentence: לכן , הממשלה מתנגדת להצעת החוק הזאת , כדי להביא לפתרון הבעיה בדרך אחרת .
DictaBERT tokens: להצעת, כדי, אחרת

Original sentence: ההצמדה [*] לא הוראת שעה .
DictaBERT sentence: ההצמדה היא כבר לא הוראת שעה .
DictaBERT tokens: היא, כבר

Original sentence: עזבי את [*] 84 , אני לא מכיר את חוק [*] והבנייה .
DictaBERT sentence: עזבי את סעיף 84 , אני לא מכיר את חוק התכנון והבנייה .
DictaBERT tokens: סעיף, התכנון

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

Original sentence: מסיבות שונות , [*] 1,275 התיקים שנפתחו בשנת [*] - טרם הוגשו כתבי אישום .
DictaBERT sentence: מסיבות שונות , בכל 1,275 התיקים שנפתחו בשנת שנפתחו - טרם הוגשו כתבי אישום .
DictaBERT tokens: בכל, שנפתחו

Original sentence: בסופו של דבר , גם אני הייתי [*] אליפויות העולם רצוף שבע שנים , וכן יצאנו עם דגל ישראל ואני יודעת שחוץ מאימא שלי , אף [*] לא יודע שבאמת אנחנו מייצגים את ישראל .
DictaBERT sentence: בסופו של דבר , גם אני הייתי בכל אליפויות העולם רצוף שבע שנים , וכן יצאנו עם דגל ישראל ואני יודעת שחוץ מאימא שלי , אף אחד לא יודע שבאמת אנחנו מייצגים את ישראל " .
DictaBERT tokens: בכל, אחד, "

Original sentence: בעצם הגענו להסכמות [*] רחבות , שקשורות בעיקר לצרכים של השקיפות בכל מה שקשור בתהליך קבלת התרומות האלה .
DictaBERT sentence: בעצם הגענו להסכמות מאוד רחבות , שקשורות בעיקר לצרכים של השקיפות בכל מה שקשור בתהליך קבלת התרומות האלה " .
DictaBERT tokens: מאוד, "

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.

2. 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:

Original sentence: אבל הנושא הוא [*] אם אתה בעד [*] ההתנתקות או נגד .
Committee sentence: אבל הנושא הוא צריך אם אתה בעד . ההתנתקות או נגד .
Committee tokens: צריך,.
Probability of committee sentence in committee corpus: -83.088
Probability of committee sentence in plenary corpus: -75.356
This sentence is more likely to appear in corpus: plenary
Plenary sentence: אבל הנושא הוא נושא אם אתה בעד או ההתנתקות או נגד .
Plenary tokens: נושא,או
Probability of plenary sentence in plenary corpus: -70.322
Probability of plenary sentence in committee corpus: -90.4
This sentence is more likely to appear in corpus: plenary

Original sentence: אבל הנושא הוא [*] אם אתה בעד [*] ההתנתקות או נגד .
DictaBERT sentence: אבל הנושא הוא לא אם אתה בעד תוכנית ההתנתקות או נגד .
DictaBERT tokens: לא, תוכנית

Other example:

```
Original sentence: עזבי את [*] 84 , אני לא מכיר את חוק [*] והבנייה .
Committee sentence:עזבי את זה 84 , אני לא מכיר את חוק חובת והבנייה .
Committee tokens: זה,חובת
Probability of committee sentence in committee corpus: -83.109
Probability of committee sentence in plenary corpus: -86.292
This sentence is more likely to appear in corpus: committee
Plenary sentence: עזבי את זה 84 , אני לא מכיר את חוק ההסדרים והבנייה .
Plenary tokens: זה,ההסדרים
Probability of plenary sentence in plenary corpus: -83.718
Probability of plenary sentence in committee corpus: -83.758
This sentence is more likely to appear in corpus: plenary
```

```
Original sentence: עזבי את [*] 84 , אני לא מכיר את חוק [*] והבנייה .
DictaBERT sentence: עזבי את סעיף 84 , אני לא מכיר את חוק התכנון והבנייה .
DictaBERT tokens: סעיף, התכנון
```

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

3. 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.

Images of the Output:

dictabert_results.txt

knesset_dictabert.py | dictabert_results.txt

```
Original sentence: יש צורך [*] ביצירת מקומות עבודה רבים יותר , בתשלום שכר [*] בעד העבודה .
DictaBERT sentence: יש צורך גם ביצירת מקומות עבודה רבים יותר , בתשלום שכר גבוה בעד העבודה .
DictaBERT tokens: גם, גבוה

Original sentence: אבל הנושא הוא [*] אם אתה בעד [*] ההתנתקות או נגד .
DictaBERT sentence: אבל הנושא הוא לא אם אתה בעד תוכנית ההתנתקות או נגד .
DictaBERT tokens: לא, תוכנית

Original sentence: לכן , הממשלה מתנגדת [*] החוק הזאת , [*] להביא לפתרון הבעיה בדרך [*] .
DictaBERT sentence: לכן , הממשלה מתנגדת להצעת החוק הזאת , כדי להביא לפתרון הבעיה בדרך אחרת .
DictaBERT tokens: להצעת, כדי, אחרת

Original sentence: ההצמדה [*] לא הוראת שעה .
DictaBERT sentence: ההצמדה היא כבר לא הוראת שעה .
DictaBERT tokens: היא, כבר

Original sentence: עדבי את [*] 84 , אני לא מכיר את חוק [*] והבנייה .
DictaBERT sentence: עדבי את סעיף 84 , אני לא מכיר את חוק התכנון והבנייה .
DictaBERT tokens: סעיף, התכנון

Original sentence: עלות החוק [*] מיליארד [*] .
DictaBERT sentence: עלות החוק היא מיליארד שקל .
DictaBERT tokens: היא, שקל

Original sentence: מסיבות שונות , [*] 1,275 התיקים שנפתחו בשנת [*] – טרם הוגשו כתבי אישום .
DictaBERT sentence: מסיבות שונות , בכל 1,275 התיקים שנפתחו בשנת שנפתחו – טרם הוגשו כתבי אישום .
DictaBERT tokens: בכל, שנפתחו

Original sentence: [*] , [*] , על תשובתך המקיפה .
DictaBERT sentence: תודה לך , מיכל , על תשובתך המקיפה .
DictaBERT tokens: תודה, מיכל

Original sentence: [*] אומר רק דבר אחד : בוא נראה את [*] האחר של המטבע .
DictaBERT sentence: אני אומר רק דבר אחד : בוא נראה את הצד האחר של המטבע .
DictaBERT tokens: אני, הצד

Original sentence: אני מנסה להבין את הפער הזה , איך יכול [*] שמצד אחד הכל [*] , הכל תחת רגולציה ומצד שני , [*] ומזהמים ברישיון .
DictaBERT sentence: אני מנסה להבין את הפער הזה , איך יכול להיות שמצד אחד הכל בסדר , הכל תחת רגולציה ומצד שני , הכל ומזהמים ברישיון .
DictaBERT tokens: להיות, בסדר, הכל

Original sentence: בסופו של דבר , גם אני הייתי [*] אליפויות העולם רצוף שבע שנים , וכן יצאנו עם דגל ישראל ואני יודעת שחוץ מאימא שלי , אף [*] לא יודע שבאמת אנחנו מייצגים את ישראל [*] .
DictaBERT sentence: בסופו של דבר , גם אני הייתי בכל אליפויות העולם רצוף שבע שנים , וכן יצאנו עם דגל ישראל ואני יודעת שחוץ מאימא שלי , אף אחד לא יודע שבאמת אנחנו מייצגים את ישראל " .
DictaBERT tokens: בכל, אחד, "

Original sentence: אבל אומרים לו , תשמע , [*] אצל [*] מאוד מאוד ארוך וכדאי שתעשה את הבדיקה [*] .
DictaBERT sentence: אבל אומרים לו , תשמע , תחור אצל הרופא מאוד מאוד ארוך וכדאי שתעשה את הבדיקה הזאת .
DictaBERT tokens: תחור, הרופא, הזאת

Original sentence: [*] באנגלית [*] העולם הם .
DictaBERT sentence: עיתוני העולם הם ברובם באנגלית .
```

knesset_dictabert.py | dictabert_results.txt

```
DictaBERT sentence: עיתוני העולם הם ברובם באנגלית .
DictaBERT tokens: עיתוני, ברובם

Original sentence: אז היא [*] אותו : [*] , אתה קיצצת ולא נתת משכורת לחורה שלי ?
DictaBERT sentence: אז היא שאלה אותו : מה , אתה קיצצת ולא נתת משכורת לחורה שלי ?
DictaBERT tokens: שאלה, מה

Original sentence: בעצם הגענו להסכמות [*] רחבות , שקשורות בעיקר לצרכים של השקיפות בכל מה שקשור בתהליך קבלת התרומות האלה [*] .
DictaBERT sentence: " בעצם הגענו להסכמות מאוד רחבות , שקשורות בעיקר לצרכים של השקיפות בכל מה שקשור בתהליך קבלת התרומות האלה .
DictaBERT tokens: ", מאוד

Original sentence: אני חושבת שזה מאוד לא נכון [*] כזה דיון שרואים שבאמת הציבור [*] בבית ורואה שאף אחד מחברי [*] לא מופיע .
DictaBERT sentence: אני חושבת שזה מאוד לא נכון לנהל כזה דיון שרואים שבאמת הציבור יושב בבית ורואה שאף אחד מחברי הכנסת לא מופיע .
DictaBERT tokens: לנהל, יושב, הכנסת

Original sentence: הרב , זה די קשה להמשיך [*] לעשות סקי ולהגיע להישגים , כי גם יש לנו [*] אחד , היחיד בארץ , שהוא נמצא בפארק דרום בתל-אביב , וגם אין מספיק מודעות לספורט הזה , במדינה [*]
DictaBERT sentence: הרב , זה די קשה להמשיך , לעשות סקי ולהגיע להישגים , כי גם יש לנו אתר אחד , היחיד בארץ , שהוא נמצא בפארק דרום בתל-אביב , וגם אין מספיק מודעות לספורט הזה , במדינה לצערי
DictaBERT tokens: ולצערי, ,, אתר

Original sentence: הוא נוסד בשנת 2002 על ידי ג'ניפר לסלו מהחזרתי ובמשך 8 שנים ניהלנו את קבוצות המוקד מלום [*] לבייג'ינג בעניין מסרים מנצחים [*] לתקשר באשר לישראל .
DictaBERT sentence: הוא נוסד בשנת 2002 על ידי ג'ניפר לסלו מהחזרתי ובמשך 8 שנים ניהלנו את קבוצות המוקד מלום קבוצות לבייג'ינג בעניין מסרים מנצחים ' לתקשר באשר לישראל .
DictaBERT tokens: קבוצות, '

Original sentence: את נושא [*] , לדוגמה .
DictaBERT sentence: בחר את נושא השיחה , לדוגמה .
DictaBERT tokens: בחר, השיחה

Original sentence: [*] [*] מראש [*] לעצור ?
DictaBERT sentence: למה לא לעצור ולחשוב מראש ?
DictaBERT tokens: למה, לא, ולחשוב

Original sentence: [*] בשום [*] זה לא
DictaBERT sentence: זה לא פוגע בשום דבר
DictaBERT tokens: פוגע, דבר

Original sentence: אנחנו צריכים [*] טובי המשפטנים .
DictaBERT sentence: אנחנו צריכים את טובי המשפטנים .
DictaBERT tokens: את

Original sentence: [*] תחום שצריך [*] התייחסות מחוקקת .
DictaBERT sentence: זה תחום שצריך לקבל התייחסות מחוקקת .
DictaBERT tokens: זה, לקבל

Original sentence: [*] הוא מאוד מוגבר [*] נושא , [*] לא אתן .
DictaBERT sentence: אני לא אתן , נושא זה הוא מאוד מוגבר .
DictaBERT tokens: אני, זה

Original sentence: [*] התיאום במסגרת הוועדה [*] מסתייע , אנחנו נעשה [*] ולא אפשרי לעשות מסתייע [*] לא אם .
DictaBERT sentence: אם לא ניתן ולא אפשרי לעשות מסתייע , אנחנו נעשה את התיאום במסגרת הוועדה .
DictaBERT tokens: את, לעשות, ניתן
```