

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 ([show results after checking the format](#)):

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:

```
def embeddings_of_sentences(sentences,model):
    try:
        sentence_dir = []
        for i,sentence in enumerate(sentences):
            tokens = make_list_of_sentence(sentence)
            avarage = np.zeros(100)
            for token in tokens:
                avarage += model.wv[token]
            avarage = avarage / len(tokens)
            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 ([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"

פותר: 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 חברים ("תודה", "חברי", "רבתי") 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 Kneset 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
רע, טוב: 0.8434505462646484
שמן, עצוב: 0.5723463296890259
עשיר, עני: 0.5578402876853943
אחר, ראשון: 0.8434258103370667
שנאה, אהבה: 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 . זה בדיוק אותו דבר .: most similar sentence: אבל זה אותו דבר
2 . את זה צריך להביא בחשבון .: most similar sentence: ולכן , צריך להביא את זה בחשבון
3 . לא , אני מוכן לבדוק את זה .: most similar sentence: אני לא מוכן לקבל את זה
4 . אם כן , אנחנו עוברים להצבעה .: most similar sentence: אם כן , רבותי , אנחנו עוברים להצבעה
5 . מבחינתך זה דבר שהוא לא נורא .: most similar sentence: זה לא דבר שהוא חדש
6 ? מה הקונספציה שלכם בעניין הזה : most similar sentence: מה התפקיד שלכם בנושא הזה
7 . אני אומר את דעתי , ודעתי שונה מדעתך .: most similar sentence: ? בגלל שאני אומר את האמת
8 . היום , בכל אופן , לא תהיה הצבעה .: most similar sentence: בכל מקרה ההצבעה לא תתקיים היום
9 . אני לא כל כך הבנתי .: most similar sentence: אני לא כל כך מבין
10 ? איך קורה דבר כזה : most similar sentence: ? איך ייתכן 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 tried 25, 100, 50 and 10 and we 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					for chunk size 3					for chunk size 5				
KNN with corss validation:					KNN with corss validation:					KNN with corss validation:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.61	0.84	0.71	29476	committee	0.69	0.88	0.77	9802	committee	0.74	0.88	0.80	5872
plenary	0.74	0.47	0.57	29476	plenary	0.83	0.60	0.70	9802	plenary	0.85	0.69	0.76	5872
accuracy			0.65	58952	accuracy			0.74	19604	accuracy			0.78	11744
macro avg	0.68	0.65	0.64	58952	macro avg	0.76	0.74	0.73	19604	macro avg	0.79	0.78	0.78	11744
weighted avg	0.68	0.65	0.64	58952	weighted avg	0.76	0.74	0.73	19604	weighted avg	0.79	0.78	0.78	11744
KNN with split:					KNN with split:					KNN with split:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.62	0.84	0.71	2948	committee	0.68	0.87	0.76	981	committee	0.72	0.87	0.79	588
plenary	0.74	0.48	0.58	2948	plenary	0.82	0.59	0.69	980	plenary	0.83	0.66	0.74	587
accuracy			0.66	5896	accuracy			0.73	1961	accuracy			0.77	1175
macro avg	0.68	0.66	0.65	5896	macro avg	0.75	0.73	0.72	1961	macro avg	0.78	0.77	0.76	1175
weighted avg	0.68	0.66	0.65	5896	weighted avg	0.75	0.73	0.72	1961	weighted avg	0.78	0.77	0.76	1175

For 50:

for chunk size 1					for chunk size 3					for chunk size 5				
KNN with corss validation:					KNN with corss validation:					KNN with corss validation:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.61	0.82	0.70	29476	committee	0.69	0.87	0.77	9802	committee	0.74	0.87	0.80	5872
plenary	0.73	0.49	0.58	29476	plenary	0.82	0.61	0.70	9802	plenary	0.85	0.70	0.76	5872
accuracy			0.65	58952	accuracy			0.74	19604	accuracy			0.78	11744
macro avg	0.67	0.65	0.64	58952	macro avg	0.76	0.74	0.73	19604	macro avg	0.79	0.78	0.78	11744
weighted avg	0.67	0.65	0.64	58952	weighted avg	0.76	0.74	0.73	19604	weighted avg	0.79	0.78	0.78	11744
KNN with split:					KNN with split:					KNN with split:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.62	0.82	0.70	2948	committee	0.68	0.85	0.76	981	committee	0.73	0.85	0.79	588
plenary	0.73	0.49	0.59	2948	plenary	0.80	0.61	0.69	980	plenary	0.82	0.69	0.75	587
accuracy			0.65	5896	accuracy			0.73	1961	accuracy			0.77	1175
macro avg	0.67	0.65	0.65	5896	macro avg	0.74	0.73	0.73	1961	macro avg	0.78	0.77	0.77	1175
weighted avg	0.67	0.65	0.65	5896	weighted avg	0.74	0.73	0.73	1961	weighted avg	0.78	0.77	0.77	1175

For 25:

for chunk size 1					for chunk size 3					for chunk size 5				
KNN with corss validation:					KNN with corss validation:					KNN with corss validation:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.62	0.77	0.69	29476	committee	0.70	0.85	0.76	9802	committee	0.75	0.85	0.80	5872
plenary	0.70	0.53	0.61	29476	plenary	0.80	0.63	0.71	9802	plenary	0.83	0.72	0.77	5872
accuracy			0.65	58952	accuracy			0.74	19604	accuracy			0.78	11744
macro avg	0.66	0.65	0.65	58952	macro avg	0.75	0.74	0.74	19604	macro avg	0.79	0.78	0.78	11744
weighted avg	0.66	0.65	0.65	58952	weighted avg	0.75	0.74	0.74	19604	weighted avg	0.79	0.78	0.78	11744
KNN with split:					KNN with split:					KNN with split:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.63	0.76	0.69	2948	committee	0.70	0.84	0.76	981	committee	0.74	0.84	0.79	588
plenary	0.70	0.54	0.61	2948	plenary	0.80	0.63	0.71	980	plenary	0.81	0.71	0.76	587
accuracy			0.65	5896	accuracy			0.74	1961	accuracy			0.77	1175
macro avg	0.66	0.65	0.65	5896	macro avg	0.75	0.74	0.74	1961	macro avg	0.78	0.77	0.77	1175
weighted avg	0.66	0.65	0.65	5896	weighted avg	0.75	0.74	0.74	1961	weighted avg	0.78	0.77	0.77	1175

For 10:

for chunk size 1					for chunk size 3					for chunk size 5				
KNN with corss validation:					KNN with corss validation:					KNN with corss validation:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.62	0.79	0.60	29476	committee	0.68	0.85	0.76	9802	committee	0.74	0.86	0.80	5872
plenary	0.71	0.51	0.59	29476	plenary	0.81	0.60	0.69	9802	plenary	0.84	0.70	0.76	5872
accuracy			0.65	58952	accuracy			0.73	19604	accuracy			0.78	11744
macro avg	0.66	0.65	0.64	58952	macro avg	0.74	0.73	0.73	19604	macro avg	0.79	0.78	0.78	11744
weighted avg	0.66	0.65	0.64	58952	weighted avg	0.74	0.73	0.73	19604	weighted avg	0.79	0.78	0.78	11744
KNN with split:					KNN with split:					KNN with split:				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
committee	0.61	0.78	0.69	2948	committee	0.68	0.85	0.76	981	committee	0.73	0.85	0.78	588
plenary	0.70	0.51	0.59	2948	plenary	0.80	0.61	0.69	980	plenary	0.82	0.68	0.74	587
accuracy			0.64	5896	accuracy			0.73	1961	accuracy			0.77	1175
macro avg	0.66	0.64	0.64	5896	macro avg	0.74	0.73	0.72	1961	macro avg	0.77	0.77	0.76	1175
weighted avg	0.66	0.64	0.64	5896	weighted avg	0.74	0.73	0.72	1961	weighted avg	0.77	0.77	0.76	1175

Questions:

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

for chunk size 1				
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
KNN 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
KNN 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 [י] כיצירת מקומות עבודה רבים יותר, בתשלום שכר [י] בעד העבודה.
DictaBERT sentence: יש צורך גם כיצירת מקומות עבודה רבים יותר, בתשלום שכר גבוה בעד העבודה.
DictaBERT tokens: גם, גבוה

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

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

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

Original sentence: עזבי את [י] 84, אני לא מכיר את חוק [י] 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, [י] 1,275 התיקים שנפתחו בשנת [י] - 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: [י] בסופו של דבר, גם אני הייתי [י] אליפויות העולם רצף שבע שנים, וכן יצאנו עם דגל ישראל ואני יודעת שחוף מאימא שלי, אף [י] לא יודע שבאמת אנחנו מייצגים את ישראל.
DictaBERT sentence: " בסופו של דבר, גם אני הייתי בכל אליפויות העולם רצף שבע שנים, וכן יצאנו עם דגל ישראל ואני יודעת שחוף מאימא שלי, אף אחד לא יודע שבאמת אנחנו מייצגים את ישראל.
DictaBERT tokens: אחד, בכל, "

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

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.

- 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 [י] אם אתה בעד [י] 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, א, .

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 [י] אם אתה בעד [י] 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:

Knesset_similar_sentences.txt:

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

Knesset_similar_words.txt:

kneset_dictabert.py kneset_similar_words.txt X

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

red_words_sentences.txt:

```
kneset_dictabert.py  red_words_sentences.txt X
1 . ברוכים הבאים , הכנסו בבקשה לחדר .: ברוכים הבאים , הכנסו בבקשה לאולם
2 . אני מזכנה 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 התבררנו שחברינו היקר לא ימשיך איתנו בשנה הבאה .: עמיתי , הערב התבררנו שחברינו הגדול לא ימשיך איתנו השנה הבאה
```

dictabert_results.txt:

```
kneset_dictabert.py dictabert_results.txt
1 . יש צורך [a] ביצירת קטגוריות עבורה רבים יותר, במהלום שכי יומר, בעד העבודה
2 DictaBERT sentence: . יש צורך גם ביצירת קטגוריות עבורה רבים יותר, במהלום שכי יומר, בעד העבודה
3 DictaBERT tokens: גם, יומר, בעד
4
5 . אבל הנושא הוא [a] אם אתה בעד [a] ההתחזקות או נגד
6 DictaBERT sentence: . אבל הנושא הוא לא אם אתה בעד חוכניות ההתחזקות או נגד
7 DictaBERT tokens: לא, חוכניות
8
9 . [a] לכן, המטלה מתנגדת [a] תחוק הזאת, [a] להביא לפתרון הבטיה בדרך
10 DictaBERT sentence: . לכן, המטלה מתנגדת להצעת תחוק הזאת, כדי להביא לפתרון הבטיה בדרך אחרת
11 DictaBERT tokens: להצעת, כדי, אחרת
12
13 . התצפדה [a] [a] לא הוראת שעה
14 DictaBERT sentence: . התצפדה היא כבר לא הוראת שעה
15 DictaBERT tokens: היא, כבר
16
17 . עזבו את [a] 84, אני לא מכיר את חוק [a] והבנויות
18 DictaBERT sentence: . עזבו את סעיף 84, אני לא מכיר את חוק התכנון והבנויות
19 DictaBERT tokens: סעיף, התכנון
20
21 . [a] עלות תחוק [a] מיליארד
22 DictaBERT sentence: . עלות תחוק היא מיליארד שקל
23 DictaBERT tokens: היא, שקל
24
25 . מסיבות שונות, [a] 1,275 תחיקים שנפתחו בשנת [a] - סרס הוגשו כחבי אישום
26 DictaBERT sentence: . מסיבות שונות, בכל 1,275 תחיקים שנפתחו בשנת שנפתחו - סרס הוגשו כחבי אישום
27 DictaBERT tokens: בכל, שנפתחו
28
29 . לד, [a], של תשובות התקופה [a]
30 DictaBERT sentence: . מורה לד, מילל, של תשובות התקופה
31 DictaBERT tokens: מורה, מילל
32
33 . אומר רק דבר אחד: בוא נראה את [a] האחר של המטבע [a]
34 DictaBERT sentence: . אני אומר רק דבר אחד: בוא נראה את הצד האחר של המטבע
35 DictaBERT tokens: אני, הצד
36
37 . אני מנסה להבין את הפטר הזה, איך יכול [a] שמעך אחד הכל [a], הכל תחת רגולציה ומעד שני, [a] ומזהמים ברישיון
38 DictaBERT sentence: . אני מנסה להבין את הפטר הזה, איך יכול להיות שמעך אחד הכל בסדר, הכל תחת רגולציה ומעד שני, הכל ומזהמים ברישיון
39 DictaBERT tokens: להיות, בסדר, הכל
40
41 . [a] בסופו של דבר, גם אני חיימ [a] אליפויות העולם רצוף שבע שנים, וכן יצאנו עם דגל ישראל ואני יודעת שחוק מתיאם שלי, אם [a] לא יודע שבאמת אנחנו מייצגים את ישראל
42 DictaBERT sentence: . " בסופו של דבר, גם אני חיימ בכל אליפויות העולם רצוף שבע שנים, וכן יצאנו עם דגל ישראל ואני יודעת שחוק מתיאם שלי, אף אחד לא יודע שבאמת אנחנו מייצגים את ישראל
43 DictaBERT tokens: " ,בכל, אחד
44
45 . [a] אבל אומרים לו, חשט, [a] אצל [a] מאוד מאוד ארוך וכדאי שתשעה את הבדיקה
46 DictaBERT sentence: . אבל אומרים לו, חשט, המור אצל הופא מאוד מאוד ארוך וכדאי שתשעה את הבדיקה הזאת
47 DictaBERT tokens: המור, הופא, הזאת
48
49 . תחולס [a] באנגלית [a]
50 DictaBERT sentence: . שיחזני תחולס הם ברובם באנגלית
51
52 . שיחזני תחולס הם ברובם באנגלית
53 DictaBERT tokens: שיחזני, ברובם
54
55 ? אז היא [a] אותו: [a], אתה קיצצת ולא נתת משכורת למורה שלי
56 DictaBERT sentence: ? אז היא שאלה אותו: מה, אתה קיצצת ולא נתת משכורת למורה שלי
57 DictaBERT tokens: שאלה, מה
58
59 . בעצם הגענו להסכמות [a] רחבות, שקשורות בעיקר לצרכים של השקיפות בכל מה שקשור בתהליך קבלת ההחלטות האלה [a]
60 DictaBERT sentence: . בעצם הגענו להסכמות מאוד רחבות, שקשורות בעיקר לצרכים של השקיפות בכל מה שקשור בתהליך קבלת ההחלטות האלה
61 DictaBERT tokens: מאוד
62
63 . אני חושבת שזה מאוד לא נכון [a] כזה דיון שרואים שבאמת תציבור [a] ביות נרואה ששך אחד מחברי [a] לא חפצים
64 DictaBERT sentence: . אני חושבת שזה מאוד לא נכון לנהל כזה דיון שרואים שבאמת תציבור יושב בבית נרואה ששך אחד מחברי הכנסת לא חפצים
65 DictaBERT tokens: לנהל, יושב, הכנסת
66
67 . [a] הרב, זה די קשה להחשיך [a] לעשות סעי' ולהגיש להישגים, כי גם יש לנו [a] אחד, היחיד בארץ, שהוא נמצא בפארק דרום בתל-אביב, וגם אין מספיק מודעות לספורט הזה, במדינה
68 DictaBERT sentence: לעשירי הרב, זה די קשה להחשיך, לעשות סעי' ולהגיש להישגים, כי גם יש לנו אחד, היחיד בארץ, שהוא נמצא בפארק דרום בתל-אביב, וגם אין מספיק מודעות לספורט הזה, במדינה
69 DictaBERT tokens: ולעשירי, אתר
70
71 . הוא נוסד בשנת 2002 על ידי ג'ניפר לסלו מחמורתי ובמשך 8 שנים ניהלנו את קבוצות המוקד מלס קבוצות לכייב'ינג בעניין מסרים מנצחים [a] להקשר באשר לישראל
72 DictaBERT sentence: . הוא נוסד בשנת 2002 על ידי ג'ניפר לסלו מחמורתי ובמשך 8 שנים ניהלנו את קבוצות המוקד מלס קבוצות לכייב'ינג בעניין מסרים מנצחים [a] להקשר באשר לישראל
73 DictaBERT tokens: קבוצות, "
74
75 . את נושא [a], לדוגמה [a]
76 DictaBERT sentence: . בחר את נושא השיחה, לדוגמה
77 DictaBERT tokens: בחר, השיחה
78
79 ? לעצור [a] חראש [a]
80 DictaBERT sentence: ? למה לא לעצור ולחשוב חראש
81 DictaBERT tokens: למה, לא, ולחשוב
82
83 . [a] זה לא [a] בשום [a]
84 DictaBERT sentence: זה לא פוטג בשום דבר
85 DictaBERT tokens: פוטג, דבר
86
87 . אנחנו צריכים [a] טובי החשפסנים
88 DictaBERT sentence: . אנחנו צריכים את טובי החשפסנים
89 DictaBERT tokens: את
90
91 . תחום שצריך [a] התייחסות מחוקקת [a]
92 DictaBERT sentence: . זה תחום שצריך לקבל התייחסות מחוקקת
93 DictaBERT tokens: זה, לקבל
94
95 . לא אתן, [a] נושא [a] הוא מאוד מורכב [a]
96 DictaBERT sentence: . אני לא אתן, נושא זה הוא מאוד מורכב
97 DictaBERT tokens: אני, זה
98
99 . אם לא [a] ולא אפשרי [a] מסריט, אנחנו נעשה [a] התיאום בתסרת הוועדה
100 DictaBERT sentence: . אם לא ניתן ולא אפשרי לעשות מסריט, אנחנו נעשה את התיאום בתסרת הוועדה
101 DictaBERT tokens: ניתן, לעשות, את
```