Deep Learning School

Физтех-Школа Прикладной математики и информатики (ФПМИ) МФТИ

Some parts of the notebook are almost the copy of [mmta-team course](). Special thanks to mmta-team for making them publicly available. [Original notebook]().

Прочитайте семинар, пожалуйста, для успешного выполнения домашнего задания. В конце ноутка напишите свой вывод. Работа без вывода оценивается ниже.

Задача поиска схожих по смыслу предложений

Мы будем ранжировать вопросы StackOverflow на основе семантического векторного представления

До этого в курсе не было речи про задачу ранжировния, поэтому введем математическую формулировку

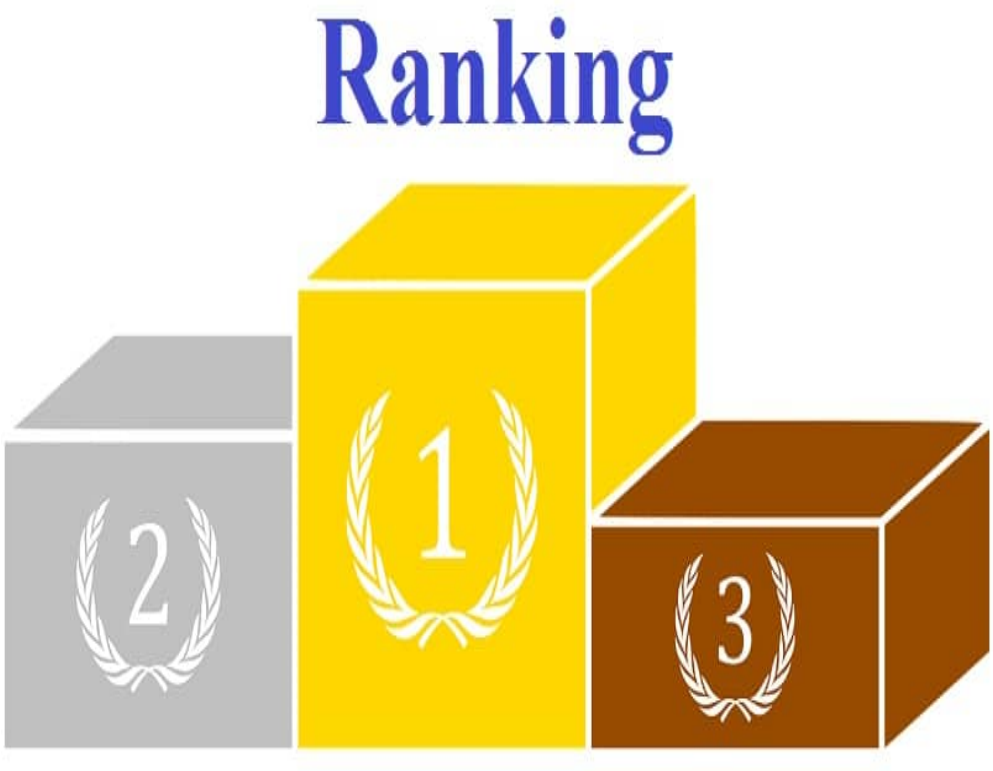
Задача ранжирования(Learning to Rank)

- X - множество объектов
- $X^l = \{x_1, x_2, \dots, x_l\}$ - обучающая выборка
  На обучающей выборке задан порядок между некоторыми элементами, то есть нам известно, что некий объект выборки более релевантный для нас, чем другой:
- $i \prec j$ - порядок пары индексов объектов на выборке X^l с индексами i и j

Задача:

построить ранжирующую функцию $a : X \to R$ такую, что

$$i \prec j \Rightarrow a(x_i) < a(x_j)$$



Embeddings

Будем использовать предобученные векторные представления слов на постах Stack Overflow.
[A word2vec model trained on Stack Overflow posts]

```
1 !wget https://zenodo.org/record/1199620/files/SO_vectors_200.bin?download=1
```

```
--2021-10-02 12:50:15--  https://zenodo.org/record/1199620/files/SO_vectors_200.bin?d
Resolving zenodo.org (zenodo.org)... 137.138.76.77
Connecting to zenodo.org (zenodo.org)|137.138.76.77|:443... connected.
```

```
HTTP request sent, awaiting response... 200 OK
Length: 1453905423 (1.4G) [application/octet-stream]
Saving to: ‘SO_vectors_200.bin?download=1.2’

SO_vectors_200.bin? 100%[===================>]   1.35G  17.7MB/s    in 83s

2021-10-02 12:51:39 (16.8 MB/s) - ‘SO_vectors_200.bin?download=1.2’ saved [145390542
```

```
from gensim.models.keyedvectors import KeyedVectors
wv_embeddings = KeyedVectors.load_word2vec_format("SO_vectors_200.bin?download=1", bina
```

Как пользоваться этими векторами?

Посмотрим на примере одного слова, что из себя представляет embedding

```
word = 'dog'
if word in wv_embeddings:
    print(wv_embeddings[word].dtype, wv_embeddings[word].shape)
```

```
float32 (200,)
```

```
print(wv_embeddings['dog'][:20])
```

```
[ 0.6851772  -1.2778991  -0.41913974  1.3623164  -3.1675398   0.09950767
  0.6402681  -1.1245339  -0.6699619  -0.6998852   0.4936771  -0.40500194
 -3.0706816  -2.2809966   0.85798043  2.7093108   0.3492745  -0.03494101
 -0.22330493  1.2290467 ]
```

```
print(f"Num of words: {len(wv_embeddings.index2word)}")
```

```
Num of words: 1787145
```

Найдем наиболее близкие слова к слову `dog`:

Вопрос 1:

- Входит ли слов `cat` топ-5 близких слов к слову `dog`? Какое место?

```
# method most_simmilar
'''your code'''
wv_embeddings.most_similar('dog')[:5]
# wv_embeddings.similarity('dog', 'cat')
```

```
[('animal', 0.8564180135726929),
 ('dogs', 0.7880867123603821),
 ('mammal', 0.7623804807662964),
 ('cats', 0.7621253728866577),
 ('animals', 0.760793924331665)]
```

Ответ: dog не входит, но входит dogs

Векторные представления текста

Перейдем от векторных представлений отдельных слов к векторным представлениям вопросов, как к **среднему** векторов всех слов в вопросе. Если для какого-то слова нет предобученного вектора, то его нужно пропустить. Если вопрос не содержит ни одного известного слова, то нужно вернуть нулевой вектор.

```
import numpy as np
import re
# you can use your tokenizer
# for example, from nltk.tokenize import WordPunctTokenizer
class MyTokenizer:
    def __init__(self):
        pass
    def tokenize(self, text):
        return re.findall('\w+', text)

tokenizer = MyTokenizer()
```

```
print(tokenizer.tokenize("C# create cookie from string and send it"))
```

```
['C', 'create', 'cookie', 'from', 'string', 'and', 'send', 'it']
```

Беда

```
def question_to_vec(question, embeddings, tokenizer, dim=200):
    """
        question: строка
        embeddings: наше векторное представление
        dim: размер любого вектора в нашем представлении

        return: векторное представление для вопроса
    """

    '''your code'''
    words = question.split()
    result = np.array([0] * dim, dtype=float)
    n_known = 0

    for word in words:
        if word in embeddings:
            result += embeddings[word]
            n_known+=1
    if n_known != 0:
        return np.array(result/n_known,dtype=float)
    else:
```

```
22         return np.array(result,dtype=float)
```

Теперь у нас есть метод для создания векторного представления любого предложения.

Вопрос 2:

- Какая третья(с индексом 2) компонента вектора предложения `I love neural networks` (округлите до 2 знаков после запятой)?

```
1 '''your code'''
2 round(question_to_vec("I love neural networks", wv_embeddings, tokenizer)[2],2)
```

```
-1.29
```

Ответ: -1.29

```
1 # Попробуем другой токенайзер
2 from nltk.tokenize import WordPunctTokenizer
3
4 nltk_tokenizer = WordPunctTokenizer()
5 round(question_to_vec("I love neural networks", wv_embeddings, nltk_tokenizer)[2],2)
```

```
-1.29
```

Оценка близости текстов

Представим, что мы используем идеальные векторные представления слов. Тогда косинусное расстояние между дублирующими предложениями должно быть меньше, чем между случайно взятыми предложениями.

Сгенерируем для каждого из N вопросов R случайных отрицательных примеров и примешаем к ним также настоящие дубликаты. Для каждого вопроса будем ранжировать с помощью нашей модели $R+1$ примеров и смотреть на позицию дубликата. Мы хотим, чтобы дубликат был первым в ранжированном списке.

Hits@K

Первой простой метрикой будет количество корректных попаданий для какого-то K:

$$\text{Hits@K} = \frac{1}{N}\sum_{i=1}^{N}[rank\_q_i^{'} \leq K],$$

- $[x < 0] \equiv \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases}$ - индикаторная функция
- q_i - i-ый вопрос
- $q_i^{'}$ - его дубликат

- $rank\_q_i^{'}$ - позиция дубликата в ранжированном списке ближайших предложений для вопроса q_i.

DCG@K

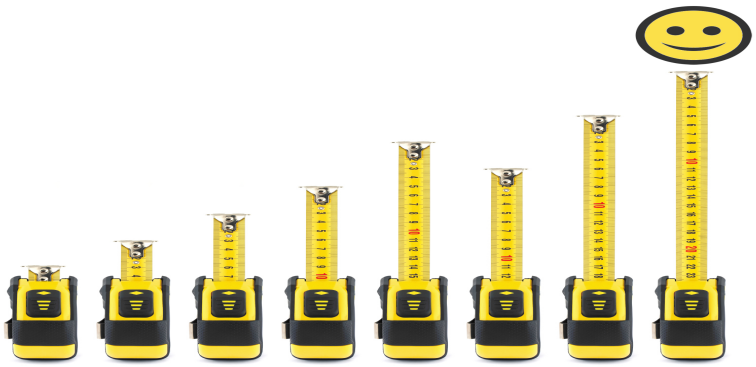
Второй метрикой будет упрощенная DCG метрика, учитывающая порядок элементов в списке путем домножения релевантности элемента на вес равный обратному логарифму номера позиции::

$$\text{DCG@K} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\log_2(1+rank\_q_i^{'})}\cdot[rank\_q_i^{'} \leq K],$$

Вопрос 3:

- Максимум `Hits@47 - DCG@1` ?

Ответ: 1 (проверял через фукции ниже)



Пример оценок

Вычислим описанные выше метрики для игрушечного примера. Пусть

- $N = 1, R = 3$
- "Что такое python?" - вопрос q_1
- "Что такое язык python?" - его дубликат $q_i^{'}$

Пусть модель выдала следующий ранжированный список кандидатов:

1. "Как изучить c++?"
2. "Что такое язык python?"
3. "Хочу учить Java"
4. "Не понимаю Tensorflow"

$\Rightarrow rank\_q_i^{'} = 2$

Вычислим метрику *Hits@K* для *K = 1, 4*:

- [K = 1] $\text{Hits@1} = [rank\_q_i^{'} \leq 1)] = 0$

- [K = 4] $\text{Hits@4} = [rank\_q_i^{'} \leq 4] = 1$

Вычислим метрику *DCG@K* для *K = 1, 4*:

- [K = 1] $\text{DCG@1} = \frac{1}{\log_2(1+2)} \cdot [2 \leq 1] = 0$
- [K = 4] $\text{DCG@4} = \frac{1}{\log_2(1+2)} \cdot [2 \leq 4] = \frac{1}{\log_2 3}$

Вопрос 4:

- Вычислите `DCG@10`, если $rank\_q_i^{'} = 9$(округлите до одного знака после запятой)

Ответ: 0.3

HITS_COUNT и DCG_SCORE

Каждая функция имеет два аргумента: dup_ranks и k. dup_ranks является списком, который содержит рейтинги дубликатов(их позиции в ранжированном списке). Например, $dup\_ranks = [2]$ для примера, описанного выше.

```
def hits_count(dup_ranks, k):
    """
        dup_ranks: list индексов дубликатов
        result: вернуть  Hits@k
    """
    N = len(dup_ranks)
    hits_value = np.sum([1 for i in range(N) if dup_ranks[i] <= k])/N
    return hits_value
```

```
def dcg_score(dup_ranks, k):
    """
        dup_ranks: list индексов дубликатов
        result: вернуть DCG@k
    """
    '''your code'''
    N = len(dup_ranks)
    dcg_value = (1/N) * sum(list(map(lambda x: int(x <= k) / np.log2(1 + x), dup_ranks)
    return dcg_value
```

Протестируем функции. Пусть $N = 1$, то есть один эксперимент. Будем искать копию вопроса и оценивать метрики.

```
import pandas as pd
```

```
copy_answers = ["How does the catch keyword determine the type of exception that was th

# наги кандидаты
```

```
candidates_ranking = [["How Can I Make These Links Rotate in PHP",
                       "How does the catch keyword determine the type of exception that
                       "NSLog array description not memory address",
                       "PECL_HTTP not recognised php ubuntu"],]
# dup_ranks — позиции наших копий, так как эксперимент один, то этот массив длины 1
dup_ranks = [candidates_ranking[i].index(copy_answers[i]) + 1 for i in range(len(copy_a

# вычисляем метрику для разных k
print('Ваш ответ HIT:', [hits_count(dup_ranks, k) for k in range(1, 5)])
print('Ваш ответ DCG:', [round(dcg_score(dup_ranks, k), 5) for k in range(1, 5)])
```

```
Ваш ответ HIT: [0.0, 1.0, 1.0, 1.0]
Ваш ответ DCG: [0.0, 0.63093, 0.63093, 0.63093]
```

У вас должно получиться

```
# correct_answers - метрика для разных k
correct_answers = pd.DataFrame([[0, 1, 1, 1], [0, 1 / (np.log2(3)), 1 / (np.log2(3)), 1
                               index=['HITS', 'DCG'], columns=range(1,5))
correct_answers
```

	1	2	3	4
HITS	0	1.00000	1.00000	1.00000
DCG	0	0.63093	0.63093	0.63093

Данные

[arxiv link]

`train.tsv` - выборка для обучения.

В каждой строке через табуляцию записаны: **<вопрос>, <похожий вопрос>**

`validation.tsv` - тестовая выборка.

В каждой строке через табуляцию записаны: **<вопрос>, <похожий вопрос>, <отрицательный пример 1>, <отрицательный пример 2>, ...**

```
from google.colab import drive
drive.mount('/content/gdrive')
!unzip /content/gdrive/MyDrive/Colab_Notebooks/simple_embeddings/stackoverflow_similar_
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive
Archive:  /content/gdrive/MyDrive/Colab_Notebooks/simple_embeddings/stackoverflow_sim
replace data/.DS_Store? [y]es, [n]o, [A]ll, [N]one, [r]ename: A
  inflating: data/.DS_Store
  inflating: __MACOSX/data/._.DS_Store
  inflating: data/train.tsv
  inflating: data/validation.tsv
```

Считайте данные.

```
def read_corpus(filename):
    data = []
    for line in open(filename, encoding='utf-8'):
        data.append(line.split('\t'))
    return data
```

Нам понадобиться только файл validation.

```
validation_data = read_corpus('./data/validation.tsv')
```

Кол-во строк

```
len(validation_data)
```

```
3760
```

```
validation_data[0]
```

```
 'How to convert the BYTEARRAYS to NSString in objective -c',
 'why is std::chrono::duration based on seconds',
 'WCF REST Client Exception',
 'Using UIBinder to create a Widget to go into a Dialog',
 'OneDrive for Business :"invalid_request","error_description":"AADSTS90014: The r
 'Are the any options for interfacing with the command prompt process from a .NET
 'Adding data in the database using microsoft access and OleDb in C#',
 'jQuery Ajax header not being passed',
 'AngularJS: Multiple views with routing without losing scope',
 'BackboneJS How to merge collections',
 'Specflow :Could not load file or assembly TechTalk.SpecFlow, Version=2.2.0.0',
 'Max byte calculated by System.Runtime.InteropServices.Marshal.SizeOf()',
 'How can my application retrieve custom fields from a DocuSign transaction?',
 'add active state to button in a button group',
 'How to define clear range for pixel color',
 'Roslyn VisualBasic.ScriptEngine doesnt recognize hostObject written on C#',
 "Neither BindingResult nor plain target object for bean name 'categoryOptions' av
 'How to change Wicket behaviour on Page Expired',
 'PHPUnit code coverage show 0% coverage',
 'How to prove code correctness lemmas with the "undefined" constant',
 'Visual Studio has insufficient privileges to debug this process. To debug this p
 'WebApi EntitySetController using composite keys',
 'How do I set up an OpenGL project using XCode 4.2 in C++?',
 'Initialize static NSString at class level',
 'Rich Text Editor inside Flux Form',
 'Setting $PATH in xampp osx',
 'How to extract data from html using PHP',
 'How to get Current location using MapModule?',
 'Excel VBA Syntax Errors & Compiling Issues',
 'How to read Logback configuration file from path outside the war file?',
 'jQuery ajax not getting every div elements',
 'TCP Socket on JAVA - Any byte >= 128 is received as 65533',
 'SqlBulkCopy keeps on throwing connection forcibly closed exception',
 'Error sending to the following VALID addresses Jenkins',
 'How to determine if a text has balanced delimiters?',
```

```
 'How to flick through a deck of cards?',
 'Selecting in SQLite Database Android',
"TSQL merge Incorrect syntax near ','",
 'Building Pocketsphinx On Android on Windows',
 'Add background image in a div without empty space between div and footer',
 'Include github project into wordpress plugin',
 'django multivaluefield & multiwidget - compress and/or decompress not working',
 'Android : Can i call my onresume() inside onpause().?',
 'How to fit picture to absolute positioning modal window?',
 'Redirect URL to PHP only if file does not exist with Nginx',
 'Laravel 4 Exception: NotFoundHttpException',
 'How To Start Using Kostache?',
 'Assigned access application exits when Ctrl + Alt + Delete is pressed.',
 'Circular dependency error when running migrations in Django 1.7c2',
 'JavaFX setOnShown fires before window is visible',
 'Flash Develop - Publishing',
 'I want to use the variable I declared somewhere else but I cannot (simple sql qu
 'buttons are not displayed',
 'Prevent Twitter responsive layout from collapsing when width < 320px',
 'Show azure cost analysis data using Azure billing API/SDK',
 'find the field names from a search query',
 'OAuth 2.0 OpenID Connect Loopback and Keycloak',
 'Windows Phone: How to retrieve the same photo from media library between applica
```

Размер нескольких первых строк

```
for i in range(5):
    print(i + 1, len(validation_data[i]))
```

```
1 1001
2 1001
3 1001
4 1001
5 1001
```

Ранжирование без обучения

Реализуйте функцию ранжирования кандидатов на основе косинусного расстояния. Функция должна по списку кандидатов вернуть отсортированный список пар (позиция в исходном списке кандидатов, кандидат). При этом позиция кандидата в полученном списке является его рейтингом (первый - лучший). Например, если исходный список кандидатов был [a, b, c], и самый похожий на исходный вопрос среди них - c, затем a, и в конце b, то функция должна вернуть список **[(2, c), (0, a), (1, b)]**.

```
from sklearn.metrics.pairwise import cosine_similarity
from copy import deepcopy
```

```
def rank_candidates(question, candidates, embeddings, tokenizer, dim=200):
    """
        question: строка
        candidates: массив строк(кандидатов) [a, b, c]
```

```
        candidates: массив строк(кандидатов) [a, b, c]
        result: пары (начальная позиция, кандидат) [(2, c), (0, a), (1, b)]
    """
    '''your code'''
    most_candidates=[]
    updated_most_candidates=[]

    q_vec=question_to_vec(question,wv_embeddings,tokenizer)
    for i in candidates:
        can_vec=question_to_vec(i,wv_embeddings,tokenizer)

        sim=cosine_similarity(can_vec.reshape(1,-1),  q_vec.reshape(1,-1))[0][0]

        most_candidates.append((sim,i))
    most_candidates.sort(key=lambda x: x[0],reverse=True)
    for i in most_candidates:
        updated_most_candidates.append((candidates.index(i[1]),i[1]))


    return updated_most_candidates
```

Протестируйте работу функции на примерах ниже. Пусть $N = 2$, то есть два эксперимента

```
questions = ['converting string to list', 'Sending array via Ajax fails']

candidates = [['Convert Google results object (pure js) to Python object', # первый экс
               'C# create cookie from string and send it',
               'How to use jQuery AJAX for an outside domain?'],

              ['Getting all list items of an unordered list in PHP',      # второй эксп
               'WPF- How to update the changes in list item of a list',
               'select2 not displaying search results']]
```

```
for question, q_candidates in zip(questions, candidates):
        ranks = rank_candidates(question, q_candidates, wv_embeddings, tokenizer)
        print(ranks, end="\n\n")
```

```
[(1, 'C# create cookie from string and send it'), (0, 'Convert Google results object

[(1, 'WPF- How to update the changes in list item of a list'), (0, 'Getting all list
```

```
for question, q_candidates in zip(questions, candidates):
        ranks = rank_candidates(question, q_candidates, wv_embeddings, nltk_tokenizer)
        print(ranks, end="\n\n")
```

```
[(1, 'C# create cookie from string and send it'), (0, 'Convert Google results object

[(1, 'WPF- How to update the changes in list item of a list'), (0, 'Getting all list
```

Для первого экперимента вы можете полностью сравнить ваши ответы и правильные ответы. Но для второго эксперимента два ответа на кандидаты будут **скрыты**(*)

```
# должно вывести
results = [[(1, 'C# create cookie from string and send it'),
            (0, 'Convert Google results object (pure js) to Python object'),
            (2, 'How to use jQuery AJAX for an outside domain?')],
           [(1, 'Getting all list items of an unordered list in PHP'), #скрыт
            (0, 'select2 not displaying search results'), #скрыт
            (2, 'WPF- How to update the changes in list item of a list')]] #скрыт
```

Последовательность начальных индексов вы должны получить `для эксперимента 1` 1, 0, 2.

Вопрос 5:

- Какую последовательность начальных индексов вы получили `для эксперимента 2` (перечисление без запятой и пробелов, например, `102` для первого эксперимента?

Ответ: 102

Теперь мы можем оценить качество нашего метода. Запустите следующие два блока кода для получения результата. Обратите внимание, что вычисление расстояния между векторами занимает некоторое время (примерно 10 минут). Можете взять для validation 1000 примеров.

```
from tqdm.notebook import tqdm
```

```
wv_ranking = []
max_validation_examples = 1000
for i, line in enumerate(tqdm(validation_data)):
    if i == max_validation_examples:
        break
    q, *ex = line
    ranks = rank_candidates(q, ex, wv_embeddings, tokenizer)
    wv_ranking.append([r[0] for r in ranks].index(0) + 1)
```

```
27%                                                1000/3760 [04:49<13:26, 3.42it/s]
```

```
for k in tqdm([1, 5, 10, 100, 500, 1000]):
    print("DCG@%4d: %.3f | Hits@%4d: %.3f" % (k, dcg_score(wv_ranking, k), k, hits_coun
```

```
100%                                                  6/6 [00:00<00:00, 65.61it/s]
DCG@   1: 0.251 | Hits@   1: 0.251
DCG@   5: 0.297 | Hits@   5: 0.339
DCG@  10: 0.314 | Hits@  10: 0.395
DCG@ 100: 0.362 | Hits@ 100: 0.626
```

Эмбеддинги, обученные на корпусе похожих вопросов

```
1 train_data = read_corpus('./data/train.tsv')
```

```
1 # Тут с лемматизацией
2
3 import spacy
4 nlp = spacy.load('en')
5
6 quest=[' '.join([str(nlp(l[0])), str(nlp(l[1]))]) for l in train_data]
7 len(quest)
```

```
1000000
```

```
1 quest=[' '.join([l[0], l[1]]) for l in train_data]
2 len(quest)
```

```
1000000
```

Улучшите качество модели.
Склеим вопросы в пары и обучим на них модель Word2Vec из gensim. Выберите размер window. Объясните свой выбор.

```
1 words_nltk = [nltk_tokenizer.tokenize(q) for q in quest]
```

```
1 words_m = [tokenizer.tokenize(q) for q in quest]
```

```
1 words_m[0]
```

```
['converting',
 'string',
 'to',
 'list',
 'Convert',
 'Google',
 'results',
 'object',
 'pure',
 'js',
 'to',
 'Python',
 'object']
```

Еще замечание: Токенайзер, который тут был удаляет знаки и некоторые слова получаются связаны

```
%%time
from gensim.models import Word2Vec
embeddings_trained = Word2Vec(words_m, # data for model to train on
                  size=200,                 # embedding vector size
                  min_count = 5,             # consider words that occured at least 5 ti
                  window=12,
                  workers=4).wv
```

```
CPU times: user 5min 8s, sys: 1.29 s, total: 5min 10s
Wall time: 2min 47s
```

```
embeddings_trained.index2word[:10]
```

```
['to', 'in', 'a', 'How', 'the', 'of', 'with', 'and', 'from', 'on']
```

```
wv_ranking = []
max_validation_examples = 1000
for i, line in enumerate(tqdm(validation_data)):
    if i == max_validation_examples:
        break
    q, *ex = line
    ranks = rank_candidates(q, ex, embeddings_trained, tokenizer)
    wv_ranking.append([r[0] for r in ranks].index(0) + 1)
```

```
27%                                                1000/3760 [04:49<12:49, 3.59it/s]
```

```
for k in tqdm([1, 5, 10, 100, 500, 1000]):
    print("DCG@%4d: %.3f | Hits@%4d: %.3f" % (k, dcg_score(wv_ranking, k), k, hits_coun
```

```
100%                                                6/6 [00:00<00:00, 86.51it/s]
DCG@   1: 0.251 | Hits@   1: 0.251
DCG@   5: 0.297 | Hits@   5: 0.339
DCG@  10: 0.314 | Hits@  10: 0.395
DCG@ 100: 0.362 | Hits@ 100: 0.626
DCG@ 500: 0.391 | Hits@ 500: 0.851
DCG@1000: 0.406 | Hits@1000: 1.000
```

1) my_tokenizer DCG@ 5: 0.427 | Hits@ 5: 0.639
2) nltk_tokenizer DCG@ 5: 0.398 | Hits@ 5: 0.542
3) default tokenizer DCG@ DCG@ 5: 0.297 | Hits@ 5: 0.339

```
# Удалим стоп слова
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
5
6 stopWords = set(stopwords.words('english'))
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
1 # Модифицированный токенайзер
2 class MyTokenizer:
3     def __init__(self):
4         pass
5     def tokenize(self, text):
6         # Не разделяем C+, C#
7         return [t for t in re.findall('\w+[+#]*', text.lower()) if t not in stopWords]
8 tokenizer = MyTokenizer()
9
10
11 words_mn = [tokenizer.tokenize(q) for q in quest]
```

Замечание:

Решить эту задачу с помощью обучения полноценной нейронной сети будет вам предложено, как часть задания в одной из домашних работ по теме "Диалоговые системы".

Напишите свой вывод о полученных результатах.

- Какой принцип токенизации даёт качество лучше и почему?
  Модификация токенайзера, который был дан, дает лучшее качество, потому что там не теряются слова по типу C++, C#, плюс убираются стоп-слова. Токенайзер из nltk работает хуже
- Помогает ли нормализация слов?
  Нормализация (лемматизация из spcy) существенного прироста качества не дала, только дольше стало происходить обработка данных
- Какие эмбеддинги лучше справляются с задачей и почему?
  Обученные эмбеддинги выдают примерно такое же качество как и предобученные, только немного лучше.
- Почему получилось плохое качество решения задачи?
  Мы обучали на парах предложений, возможно, обучение на полноценных данных дало бы качество лучше. Возможно, использование других параметров для эмбеддинга, даст качество лучше
- Предложите свой подход к решению задачи.
  Например предобученный эмбеддинг дообучить на наших данных, улучшить способ отбора нужных слов.

Вывод:

Double-click (or enter) to edit

✓ 0s completed at 4:10 PM