Упражнения по тематическому моделированию и api.vk

самостоятельное прорешивание.
ответы сразу после упражнения

Задание 1. Выкачать посты из паблика ВК (https://vk.com/meduzaproject)

Решение 1

```
1. groupname='meduzaproject'
2. url = "https://api.vk.com/method/wall.get?v=5.29&lang=en&scope=super&domain="+str(groupname)
3. first_try = json.loads(requests.get(url).text)
#делаем первую попытку выкачки постов
4. count_posts = int(first_try['response']['count'])
#выстаскиваем из этой первой попытки общее количество постов паблика
5. texts = []
6. while count > len(texts):
7.     second_try = json.loads(requests.get(url+"&offset="+str(len(texts))).text)
#через offset сдвигаем выборку постов
8.     for i in range(len(second_try['response']['items'])):
9.         s=second_try['response']['items'][i]['text']
10.        texts.append(s)
11. print texts[1]
```

Задание 2. Провести тематическое моделирование постов паблика ВК (https://vk.com/meduzaproject)

Решение 2

1. Взять за основу код из решения (1).
2. Добавить функцию лемматизации.

```
def preprocess(text):
    wnl = nltk.WordNetLemmatizer()
    return [wnl.lemmatize(t) for t in text.lower().split()]
```

3. Вместо строчки texts.append(s) написать texts.append(preprocess(s)).
4. Создать словарь.

```
dictionary = corpora.Dictionary(texts)
```

5. Отфильтровать словарь.

```
dictionary.filter_extremes(no_below = 5, no_above = 0.5)
```

6. Создать корпус.

```
corpus = [dictionary.doc2bow(text) for text in texts]
```

7. Запустить модель LDA.

```
model = models.ldamodel.LdaModel(corpus, id2word=dictionary, num_topics=5,
chunksize=50, update_every=1, passes=4)
```

8. Вывести на экран топ-слова тематик.

```
for position in range(10):
    for topic in range(0, 5):
        print model.show_topic(topic)[position][1].center(20, ' '),
    print
```
