Домашее задание 2 (к 25/02/2016). Модель Word2vec

Составьте достаточно большую коллекцию текстов на любом языке или используйте тексты отсюда: `https://www.kaggle.com/c/word2vec-nlp-tutorial/data`.

1. **[1 балла]** Обучите модель word2vec. Оцените время обучения модели, используя модуль time.

   Есть два варианта обучения модели: по отзывам целиком и с учетам границ предложений. В принципе, погрешностью, которая возникает в первом случае можно пренебречь, но если вы хотите учитывать границы предложений, то можно использовать `sent_tokenize` из `nltk.tokenize`.

2. **[2 балла]** Приведите 5-10 примеров использования `.most_similar` для определения близких слов. Корректно ли они найдены? Являются ли синонимами исходного слова?

3. **[2 балла]** Приведите 5-10 примеров использования `.most_similar` для определения ассоциаций (А к Б, как В к?). Корректно ли найдены ассоциации?

4. **[2 балла]** Приведите 5-10 примеров использования `.doesnt_match` для определения лишнего слова. Корректно ли найдены лишние слова?

5. **[3 балла]** Попробуйте найти такие пары и тройки слов, для которых

   - **не** выполняются свойства коммутативности и транзитивности относительно операции определения близких слов.
   - выполняются свойства коммутативности и транзитивности относительно операции определения близких слов.

   Обозначим отношение "входить в топ-3 по `.most_similar`" символом $\circ$.

   Коммутативность $x \circ y \implies y \circ x$

   Транзитивность $x \circ y, y \circ z \implies x \circ z$