tf-idf

tf-idfとは**tf**と**idf**から文章中の単語の重要度を算出する手法である

$$\text{tf-idf} = tf \times idf$$

tf

tf (Term Frequency) は単語の頻出度を表す
文書中に出現する頻度が多いほど、重要な単語である可能性が高い

$$tf(t, d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}}$$

$tf(t, d)$: 文書 d 内のある単語 t のTF値

$n_{t,d}$: ある単語 t の文書 d 内での出現回数

$\sum_{s \in d} n_{s,d}$: 文書 d 内のすべての単語の出現回数の和

idf

idf (Inverse Document Frequency) はその単語がいくつの文章で共通して使われているかを表す
多くの文章で使われているほどその単語の重要度は低い

$$\text{idf}(t) = \log \frac{N}{df(t)} + 1$$

※対数の底はなんでもよい(今回は自然対数とする)

$idf(t)$: ある単語 t のIDF値

N : 全文書数

$df(t)$: ある単語 t が出現する文書の数

例

文章１:私はリンゴとリンゴが好きです。
文章２:私はリンゴとミカンが好きです。
文章３:私は虫が嫌いです。

以上のような３つの文章があるとします。

1. 形態素解析
   このままではtf-idfを計算できないので Mecab などを用いて形態素解析を行います。(今回は名詞のみを抜き取ります)
   私 / は / リンゴ / と / リンゴ / が / 大好き / です / 。→　私,リンゴ,リンゴ
   私 / は / リンゴ / と / ミカン / が / 好き / です / 。→　私,リンゴ,ミカン
   私 / は / 虫 / が / 嫌い / です / 。→　私,虫
   私,**リンゴ**,**ミカン**,**虫**
   の4つの名詞から構成されていることがわかります。
2. tfの計算
   $\mathrm{tf}(\quad 私 \quad, 文章１) = \frac{1}{3}$
   $\mathrm{tf}(リンゴ, 文章１) = \frac{2}{3}$
   $\mathrm{tf}(\quad 私 \quad, 文章２) = \frac{1}{3}$
   $\mathrm{tf}(リンゴ, 文章２) = \frac{1}{3}$
   $\mathrm{tf}(みかん, 文章２) = \frac{1}{3}$
   $\mathrm{tf}(\quad 私 \quad, 文章３) = \frac{1}{2}$
   $\mathrm{tf}(\quad 虫 \quad, 文章３) = \frac{1}{2}$
3. idfの計算
   $\mathrm{idf}(\quad 私 \quad) = \log \frac{3}{3} + 1$
   $\mathrm{idf}(リンゴ) = \log \frac{3}{2} + 1$
   $\mathrm{idf}(ミカン) = \log \frac{3}{1} + 1$
   $\mathrm{idf}(\quad 虫 \quad) = \log \frac{3}{1} + 1$
4. tf-idfの計算
   $\mathrm{tf}(\quad 私 \quad, 文章１) \times idf(\quad 私 \quad) = \frac{1}{3} \times (\log \frac{3}{3} + 1) = 0.33333...$
   $\mathrm{tf}(リンゴ, 文章１) \times idf(リンゴ) = \frac{1}{3} \times (\log \frac{3}{3} + 1) = 0.93698...$
   $\mathrm{tf}(\quad 私 \quad, 文章１) \times idf(\quad 私 \quad) = \frac{1}{3} \times (\log \frac{3}{3} + 1) = 0.33333...$
   $\mathrm{tf}(リンゴ, 文章１) \times idf(リンゴ) = \frac{1}{3} \times (\log \frac{3}{3} + 1) = 0.46849...$
   $\mathrm{tf}(ミカン, 文章１) \times idf(ミカン) = \frac{1}{3} \times (\log \frac{3}{3} + 1) = 0.69954...$
   $\mathrm{tf}(\quad 私 \quad, 文章１) \times idf(\quad 私 \quad) = \frac{1}{3} \times (\log \frac{3}{3} + 1) = 0.5$
   $\mathrm{tf}(\quad 虫 \quad, 文章１) \times idf(\quad 虫 \quad) = \frac{1}{3} \times (\log \frac{3}{3} + 1) = 1.04931....$

課題

tf , idf , tf-idf を求めるプログラムを作成してください。
以下の条件を守ってください。

- 例にあげた３つの文章を用いること
- Mecabを用いて名詞を抜き出すこと
- tf-idfを求められるようなライブラリの使用は不可（numpy,pandas等は可)
  表示等の方法はおまかせします。