

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)

$\text{idf}(t)$: 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を食べたい。

文章2:私はリンゴとミカンを食べたい。

文章3:私は虫を食べたくない。

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 / を / 食べ / たい / 。 → 私, リンゴ, リンゴ

私 / は / リンゴ / と / ミカン / を / 食べ / たい / 。 → 私, リンゴ, ミカン

私 / は / 虫 / を / 食べ / たく / ない / 。 → 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

$$\text{tf}(\text{私}, \text{文章1}) = \frac{1}{3}$$

$$\text{tf}(\text{リンゴ}, \text{文章1}) = \frac{2}{3}$$

$$\text{tf}(\text{私}, \text{文章2}) = \frac{1}{3}$$

$$\text{tf}(\text{リンゴ}, \text{文章2}) = \frac{1}{3}$$

$$\text{tf}(\text{みかん}, \text{文章2}) = \frac{1}{3}$$

$$\text{tf}(\text{私}, \text{文章3}) = \frac{1}{2}$$

$$\text{tf}(\text{虫}, \text{文章3}) = \frac{1}{2}$$

3. idfの計算

$$\text{idf}(\text{私}) = \log \frac{3}{2} + 1$$

$$\text{idf}(\text{リンゴ}) = \log \frac{3}{2} + 1$$

$$\text{idf}(\text{ミカン}) = \log \frac{3}{1} + 1$$

$$\text{idf}(\text{虫}) = \log \frac{3}{1} + 1$$

4. tf-idfの計算

$$\text{tf}(\text{私}, \text{文章1}) \times \text{idf}(\text{私}) = \frac{1}{3} \times (\log \frac{3}{2} + 1) = 0.33333...$$

$$\text{tf}(\text{リンゴ}, \text{文章1}) \times \text{idf}(\text{リンゴ}) = \frac{1}{3} \times (\log \frac{3}{2} + 1) = 0.93698...$$

$$\text{tf}(\text{私}, \text{文章1}) \times \text{idf}(\text{私}) = \frac{1}{3} \times (\log \frac{3}{2} + 1) = 0.33333...$$

$$\text{tf}(\text{リンゴ}, \text{文章1}) \times \text{idf}(\text{リンゴ}) = \frac{1}{3} \times (\log \frac{3}{2} + 1) = 0.46849...$$

$$\text{tf}(\text{ミカン}, \text{文章1}) \times \text{idf}(\text{ミカン}) = \frac{1}{3} \times (\log \frac{3}{1} + 1) = 0.69954...$$

$$\text{tf}(\text{私}, \text{文章1}) \times \text{idf}(\text{私}) = \frac{1}{3} \times (\log \frac{3}{2} + 1) = 0.5$$

$$\text{tf}(\text{虫}, \text{文章1}) \times \text{idf}(\text{虫}) = \frac{1}{3} \times (\log \frac{3}{1} + 1) = 1.04931....$$

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。