TF-IDFによる類似文章の判断

2-3-48 山本雄希

やりたいこと

- 類似文章を判断し、

  一定以上似ている文章を削除する。

  - youtubeや、twitchなど、配信サイトでのコメント欄で、同じコメントを連投された時にそれを削除する。

手法など

- phpで簡易的にコメントページを作成。
- SQliteでコメントをDBに格納。
- pythonでTF-IDFを計算後、cos類似度を求めて一定以上のコメントをDBから削除。

TF-IDF

- DBの全コメントを単語ごとに辞書にして、一つのコメントにどの単語がどの程度出現するかを行列にする。
- 最新のコメントと既存のコメント一つずつ、IF-IDFのCOS類似度を求め、一定以上のコメントを削除。

```
[[1.        0.        0.        0.        0.        0.       ]
 [0.        0.4472136 0.4472136 0.4472136 0.4472136 0.4472136]]
{'こんにちは': 0, 'はじめまして': 1, 'よろしく': 2, 'お願い': 3, 'し': 4, 'ます': 5}
['こんにちは', 'はじめましてよろしくお願いします']
```

TF-IDF

こんにちは

はじめましてよろしくお願いします

今何をしていますか

早く始めよう

```
[[1.91629073 0.         0.         0.         0.         0.
  0.         0.         0.         0.         0.         0.
  0.         0.         0.        ]
 [0.         1.91629073 1.91629073 1.91629073 1.51082562 1.51082562
  0.         0.         0.         0.         0.         0.
  0.         0.         0.        ]
 [0.         0.         0.         0.         1.51082562 1.51082562
  1.91629073 1.91629073 1.91629073 1.91629073 1.91629073 1.91629073
  0.         0.         0.        ]
 [0.         0.         0.         0.         0.         0.
  0.         0.         0.         0.         0.         0.
  1.91629073 1.91629073 1.91629073]]
```

cos類似度

- 一つのコメントに対するTF-IDFをベクトルとして考えて、最新のコメントに類似するコメントがあった場合最新のコメントを削除する。

こんにちは

はじめましてよろしくお願いします

今何をしていますか

早く始めよう

はじめましてこんにちは

```
0.7071067811865476
0.28694451355518747
0.0
0.0
```

cos類似度

- 最後に、基準とした値を超えたコメントを削除して更新する。

デモ

- はじめましてこんにちは
- 時間後どれくらい
- 早く始めろ

質疑応答

感想、考察

- コメント数や単語数が増えた時に問題が生じる可能性がある。
  - コメント数が増えると「が、を、こと、は、です、ます」などで値が大きくなりやすい。ー＞基準値を上げて対策。

- php上でpythonを動かしてターミナルを起動せずに実演したかった。（方法を模索したができなかった。）