プログラムに文章を読ませる

TD-IDF法による重要単語の検出

2022.3.22 青森大学 ソフトウェア情報学部情報学部 入学前勉強会
講師：藤澤 日明

自己紹介

本日の先生役：藤澤 日明（ふじさわ　あきら）

出身：兵庫県　姫路市

趣味：漫画、ゲーム、Vtuber

研究のテーマ・キーワード

機械学習、**データ解析**、**自然言語処理** など
「**マルチメディア**に対する情報処理や、顔文字のような**非言語表現**に関する研究を行っています」

顔文字の影響について

問1.次の文章の感情極性を判定せよ

今日のテスト終わった

ポジティブ/嬉しい ⟷ **ネガティブ/悲しい**

顔文字の影響について

問2.次の文章の感情極性を判定せよ

A: **今日のテスト終わった（^0^）**

B: **今日のテスト終わった（T^T）**

顔文字の影響について

問2.次の文章の感情極性を判定せよ

A: **今日のテスト終わった (^0^)**

B: **今日のテスト終わった (T^T)**

顔文字の表情のおかげで、文章全体の印象が変化している(固定化されている)

顔文字の影響について

問3.次の文章について、感情の度合いを順位付けせよ

A: 怒ってないよ

B: 怒ってないよ(^^)

C: 怒ってないよ(# ゜Д゜)

顔文字の影響について

問3.次の文章について、感情の度合いを順位付けせよ

A: **怒ってないよ**

B: **怒ってないよ(^^)**

C: **怒ってないよ(# ゜Д゜)**

顔文字の有無・種類によって、文章全体の印象が強調されたり、和らいだりしている

自然言語処理（コンピュータで文章を処理する）

- 自然言語処理とは『私たちが普段使っている言葉(=**自然言語**)を、コンピュータに理解させる技術・研究』のこと

  - 「この文章はどんな内容の話題を話しているのだろう」
  - 「この単語はどんな意味・目的で使われているのだろう」
  - 「この文章はポジティブ・ネガティブのどちらの感情をもつのだろう」

    など、目的は様々

自然言語処理の目的をもう少し紹介

- 単語切り分け
  - 単語の切れ目を推定する
- 形態素解析
  - (名詞、動詞、形容詞などの品詞別に分類)
- 構文解析
- 機械翻訳
- 同義語判定、略語判定
  - (言葉同士が同じ意味を持つのか判定
    『「**青大**」≒「**青森大学**」≠「**青学**」』など)
- **文章要約**

重要な単語を見つけ出す技術

- 人間でもコンピュータでも、長い文章を読むのは中々に大変
  - 特に自分の知らない分野の説明を理解することは難しい……

- 長文を理解するために必要なのは文章要約の力
  - 文章の中から重要な単語を見つけ出し、ピックアップしたり
    重複する文章を削除して、文章全体をスリムにする技術

- その文章において重要なキーワードを見つける技術は需要が高い

長文の例

SNSは世界中の人々と気軽にコミュニケーションがとれる優れたツールだ。SNSを利用した商売も行われおり、仕事でもプライベートでも欠かせない存在と言える。一方、情報リテラシーが低い一部の人間によって、SNSが社会問題のきっかけになることもある。私達はSNSをどのように利用していくべきだろうか？

私はSNSを個人を発信する場として、若いうちから積極的に利用していくべきだと考える。なぜなら、これからは「個人」が活躍しやすい社会になるからだ。一昔の日本であれば、学校を卒業したら会社に入って働くのが当たり前だった。しかし今では、クラウドソーシングを利用して仕事をしたり、起業したりする若者が増えてきている。このように個人が活躍しやすくなると、SNSが自分の名刺代わりになる。個人で活躍していく環境を作るためにも、SNSは積極的に利用するべきだ。

しかし「SNSの利用には情報リテラシーが求められる。若者には使わせるべきでない」という意見もあるだろう。たしかに、SNSで犯罪行為を発信してバッシングを受けた例は枚挙に暇がない。この問題を解決するためには、利用を制限するのではなく、義務教育で情報リテラシーに関する授業時間を増やしていくべきだろう。利用の制限をすれば、SNSとの接し方がわからない大人が育つだけだ。問題があれば何でもかんでも制限するのではなく、適切な接し方を学ばせるのが正しい対処法だろう。　　（*https://businessbook-lasdream.com/2019/07/16/paper-rei-19/*より引用）

長文の例

SNSは世界中の人々と気軽にコミュニケーションがとれる優れたツールだ。SNSを利用した商売も行われおり、仕事でもプライベートでも欠かせない存在と言える。一方、情報リテラシーが低い一部の人間によって、SNSが社会問題のきっかけになることもある。私達はSNSをどのように利用していくべきだろうか？

私はSNSを個人を発信する場として、若いうちから積極的に利用していくべきだと考える。なぜなら、これからは「個人」が活躍しやすい社会になるからだ。一昔の日本であれば、学校を卒業したら会社に入って働くのが当たり前だった。しかし今では、クラウドソーシングを利用して仕事をしたり、起業したりする若者が増えてきている。このように個人が活躍しやすくなると、SNSが自分の名刺代わりになる。個人で活躍していく環境を作るためにも、SNSは積極的に利用するべきだ。

しかし「SNSの利用には情報リテラシーが求められる。若者には使わせるべきでない」という意見もあるだろう。たしかに、SNSで犯罪行為を発信してバッシングを受けた例は枚挙に暇がない。この問題を解決するためには、利用を制限するのではなく、義務教育で情報リテラシーに関する授業時間を増やしていくべきだろう。利用の制限をすれば、SNSとの接し方がわからない大人が育つだけだ。問題があれば何でもかんでも制限するのではなく、適切な接し方を学ばせるのが正しい対処法だろう。　　*(https://businessbook-lasdream.com/2019/07/16/paper-rei-19/より引用)*

長文の例

SNSは世界中の人々と気軽にコミュニケーションがとれる優れたツールだ。SNSを利用した商売も行われおり、仕事でもプライベートでも欠かせない存在と言える。一方、情報リテラシーが低い一部の人間によって、SNSが社会問題のきっかけになることもある。私達はSNSをどのように利用していくべきだろうか？

私はSNSを個人を発信する場として、若いうちから積極的に利用していくべきだと考える。なぜなら、これからは「個人」が活躍しやすい社会になるからだ。一昔

た。しかし今では、クラウドソーシングを利用し

人が活躍しやすくなると、SNSが自分の名刺代わり

するべきだ。

しかし「SNSの利用には情報リテラシーが求

しかに、SNSで犯罪行為を発信してバッシングを受

るのではなく、義務教育で情報リテラシーに関す

し方がわからない大人が育つだけだ。問題があれ

しい対処法だろう。　　(*https://businessbook-lasdream.com/2019/07/16/paper-rei-19/より引用*)

『SNS』『利用』『情報リテラシー』
といった重要ワードをどのように検出・活用すればよいのか？

→ TF-IDF法を使ってみる

TF－IDF法

- いくつかの文章(ドキュメント)において出現する単語群について、それらの文章を特徴づけるような重要な単語を見つけるためのアルゴリズム

- TF値とIDF値の積によって得られる値

- 要素として「単語の出現頻度」「単語のレア度」を数値化している
(文章をベクトルとして扱うためのアルゴリズムでもある)

TFについて

- TF = “*Term Frequency*”(単語頻度)の略
- ある文章における、単語の出現頻度を表す

- 式としては主にTF(w_i, d_i)と書いたりする

  TF(w_i, d_i) → 文章(document)*i*における、単語(word)*i*のTF値

  ※iには単語がそのまま入ったり、単語の順番が入ったりする

TF(w_i, d_i)　＝

文章d_iにおける単語w_iの出現回数÷文章d_iに出現する単語の総出現数

TFの取得例 1/3　単語のピックアップ

文章①

SNSは世界中の人々と気軽にコミュニケーションがとれる優れたツールだ。SNSを利用した商売も行われおり、仕事でもプライベートでも欠かせない存在と言える。一方、情報リテラシーが低い一部の人間によって、SNSが社会問題のきっかけになることもある。私達はSNSをどのように利用していくべきだろうか？

文字が多いので、ひとまず候補の単語を抽出してみる

SNSは世界中の**人々**と気軽に**コミュニケーション**がとれる優れた**ツール**だ。**SNS**を利用した**商売**も行われおり、**仕事**でも**プライベート**でも欠かせない存在と言える。一方、**情報リテラシー**が低い一部の人間によって、**SNS**が**社会問題**のきっかけになることもある。私達は**SNS**をどのように利用していくべきだろうか？

TFの取得例　2/3 単語の出現数確認

文章①

SNSは世界中の人々と気軽にコミュニケーションがとれる優れたツールだ。SNSを利用した商売も行われおり、仕事でもプライベートでも欠かせない存在と言える。一方、情報リテラシーが低い一部の人間によって、SNSが社会問題のきっかけになることもある。私達はSNSをどのように利用していくべきだろうか？

それぞれの単語の出現数を表にまとめてみる

単語	SNS	人々	コミュニケーション	ツール	商売	仕事	プライベート	情報リテラシー	社会問題	合計
出現回数										

TFの取得例　2/3 単語の出現数確認

文章①

SNSは世界中の人々と気軽にコミュニケーションがとれる優れたツールだ。SNSを利用した商売も行われおり、仕事でもプライベートでも欠かせない存在と言える。一方、情報リテラシーが低い一部の人間によって、SNSが社会問題のきっかけになることもある。私達はSNSをどのように利用していくべきだろうか？

単語	SNS	人々	コミュニケーション	ツール	商売	仕事	プライベート	情報リテラシー	社会問題	合計
出現回数	4	1	1	1	1	1	1	1	1	12

TFの取得例　3/3 TFを計算する

単語	SNS	人々	コミュニケーション	ツール	商売	仕事	プライベート	情報リテラシー	社会問題	合計
出現回数	4	1	1	1	1	1	1	1	1	12

TF(SNS, 文章①) ＝ ？

TF(情報リテラシー, 文章①) ＝ ？

$TF(w_i, d_i)$ ＝
文章d_iにおける単語w_iの出現回数÷文章d_iに出現する単語の総出現数

TFの取得例　3/3 TFを計算する

単語	SNS	人々	コミュニケーション	ツール	商売	仕事	プライベート	情報リテラシー	社会問題	合計
出現回数	4	1	1	1	1	1	1	1	1	12

TF(SNS, 文章①) ＝ **4** / **12** = **0.3**

TF(情報リテラシー, 文章①) ＝ **1** / **12** = **0.08**

文章①において、「情報リテラシー」よりも「SNS」の方が重要度が高いと判断できる

例題

次の文章A-Cについて、TF（ラーメン）、TF（好き）、TF（麺）をそれぞれの文章毎に求めよ。

文章A ＝　　{ラーメン、醤油、ラーメン、スープ、コク、麺、好き}

文章B ＝　　{味噌、ラーメン、味噌、濃い、麺、太い}

文章C ＝　　{ラーメン、塩、さっぱり、好き、麺、細い、好き}

例題 解答

次の文章A-Cについて、TF（ラーメン）、TF（好き）、TF（麺）をそれぞれの文章毎に求めよ。

TF（ラーメン、文章A）＝　2 / 7
TF（好き、文章A）＝　1 / 7
TF(麺、文章A)　＝　1 / 7

TF（ラーメン、文章B）＝　1 / 6
TF（好き、文章B）＝　0 / 6
TF(麺、文章B)　＝　1 / 6

TF（ラーメン、文章C）＝　1 / 7
TF（好き、文章C）＝　2 / 7
TF(麺、文章C)　＝　1 / 7

- 同じ単語であっても、文章や単語数が異なれば当然TF値は変化する
- さらに複数の文章間で比較することで、重要単語の推測を行うことが可能になる

ところで……

本当に出現頻度が多いだけで、文章の特徴語を決定してよいのか?

TFだけでは不十分

次の三つの文章で、もっともTF値が高そうな単語は何か?

『私はラーメンが好きだ。特に私の一番好きな味は塩味で、あっさりしたスープがたまらない。』

『私はいつもお昼に中華料理屋へ行く。そこでいつも私が頼むのは特製の塩ラーメンだ。』

『私は大食いなので大盛りを注文するが、私の友達はいつも小盛りで頼んでいる。』

TFだけでは不十分

次の三つの文章で、もっともTF値が高そうな単語は何か?

『私はラーメンが好きだ。特に私の一番好きな味は塩味で、あっさりしたスープがたまらない。』

『私はいつもお昼に中華料理屋へ行く。そこでいつも私が頼むのは特製の塩ラーメンだ。』

『私は大食いなので大盛りを注文するが、私の友達はいつも小盛りを頼んでいる。』

→ 『私』という単語が頻出している ≒ 重要?

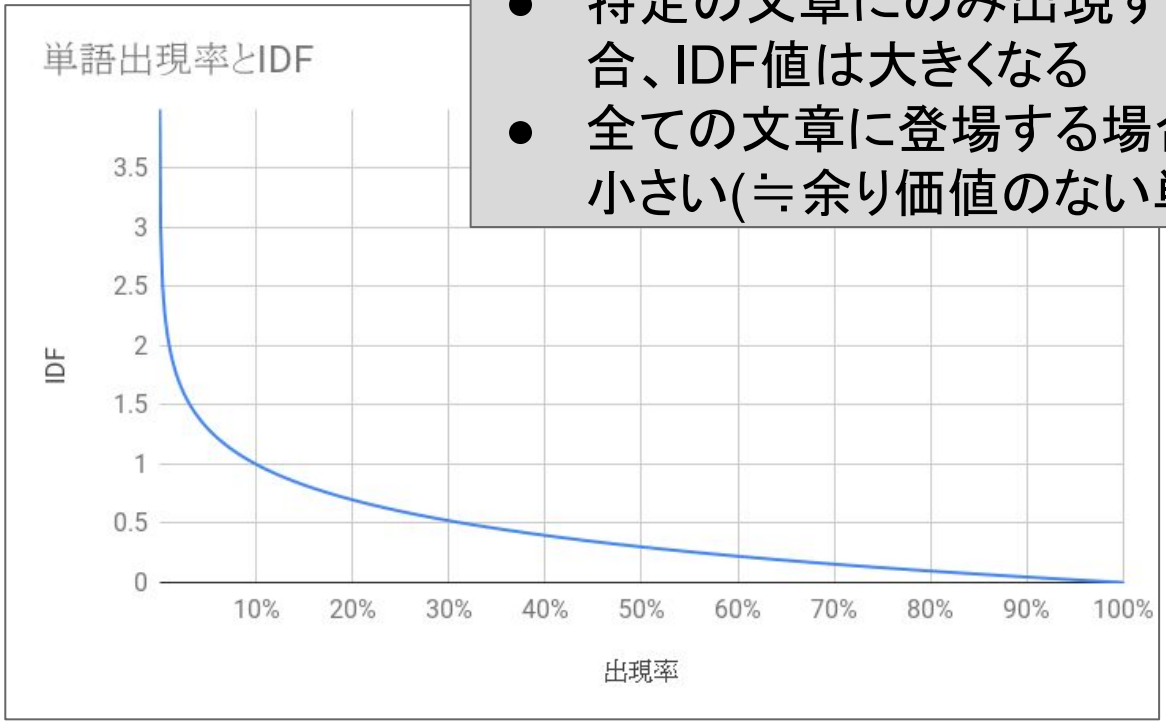
IDFについて

- IDF = “*Inverse Document Frequency*”(逆文章頻度)の略
- いくつかの文章における、注目単語の出現頻度の低さ(≒特定の文章にしか出現しないレアな単語)を表す

- 式としては主にIDF(w_i)と書いたりする

$$IDF(w_i) = \log(\text{文章の総数} \div \text{単語}w_i\text{が出現する文章の総数}) + 1$$

IDF値のイメージ



- 特定の文章にのみ出現する単語の場合、IDF値は大きくなる
- 全ての文章に登場する場合、IDF値は小さい(≒余り価値のない単語と判定)

※https://joho-ka.mints.ne.jp/tf-idf02?doing_wp_cron=1647594159.1608500480651855468750 より

IDFの例題

次の三つの文章における、IDF(私)、IDF(ラーメン)、IDF(あっさり)を計算してみる

文章A:『私、ラーメン、私、塩、あっさり、スープ』

文章B:『私、お昼、中華料理屋、私、特製、塩、ラーメン』

文章C:『私、大食い、大盛り、私、友達、小盛り』

$$IDF(w_i) = \log(\text{文章の総数} \div \text{単語}w_i\text{が出現する文章の総数}) + 1$$

IDFの例題

次の三つの文章における、IDF(私)、IDF(ラーメン)、IDF(あっさり)を計算してみる

文章A:『私、ラーメン、私、塩、あっさり、スープ』

文章B:『私、お昼、中華料理屋、私、特製、塩、ラーメン』

文章C:『私、大食い、大盛り、私、友達、小盛り』

IDF(私)
＝ log(3/3)+1 = **1**

IDF(ラーメン)
= log(3/2)+1 ≒**1.17**

IDF(あっさり)
= log(3/1)+1 ≒ **1.47**

$$IDF(w_i) = \log(\text{文章の総数} \div \text{単語}w_i\text{が出現する文章の総数}) + 1$$

最後にTF-IDFを計算してみよう

- TF-IDFはTF値とIDF値の積である
- 例題
  - 以下の文章について、「私」「ラーメン」「塩」「スープ」の単語について
    それぞれの文章におけるTF-IDF値を求めよ

文章A:『私、ラーメン、私、塩、あっさり、スープ』

文章B:『私、お昼、中華料理屋、私、特製、塩、ラーメン』

文章C:『私、大食い、大盛り、ラーメン、私、友達、小盛り、ラーメン』

TF値、IDF値は表に記すと分かりやすい

TF値

	私	ラーメン	塩	スープ
文章A				
文章B				
文章C				

IDF値

私	ラーメン	塩	スープ

TF値、IDF値は表に記すと分かりやすい

TF値

	私	ラーメン	塩	スープ
文章A	2/6	1/6	1/6	1/6
文章B	2/7	1/7	1/7	0
文章C	2/7	2/7	0	0

IDF値

私	ラーメン	塩	スープ
log(3/3)+1	log(3/2)+1	log(3/2)+1	log(3/1)+1

解答：それぞれの文章でのTF-IDF値

※実際の計算はエクセルで楽します

TF

	私	ラーメン	塩	スープ
文章A	0.3333	0.1667	0.1667	0.1667
文章B	0.2857	0.1429	0.1429	0
文章C	0.2857	0.1429	0	0

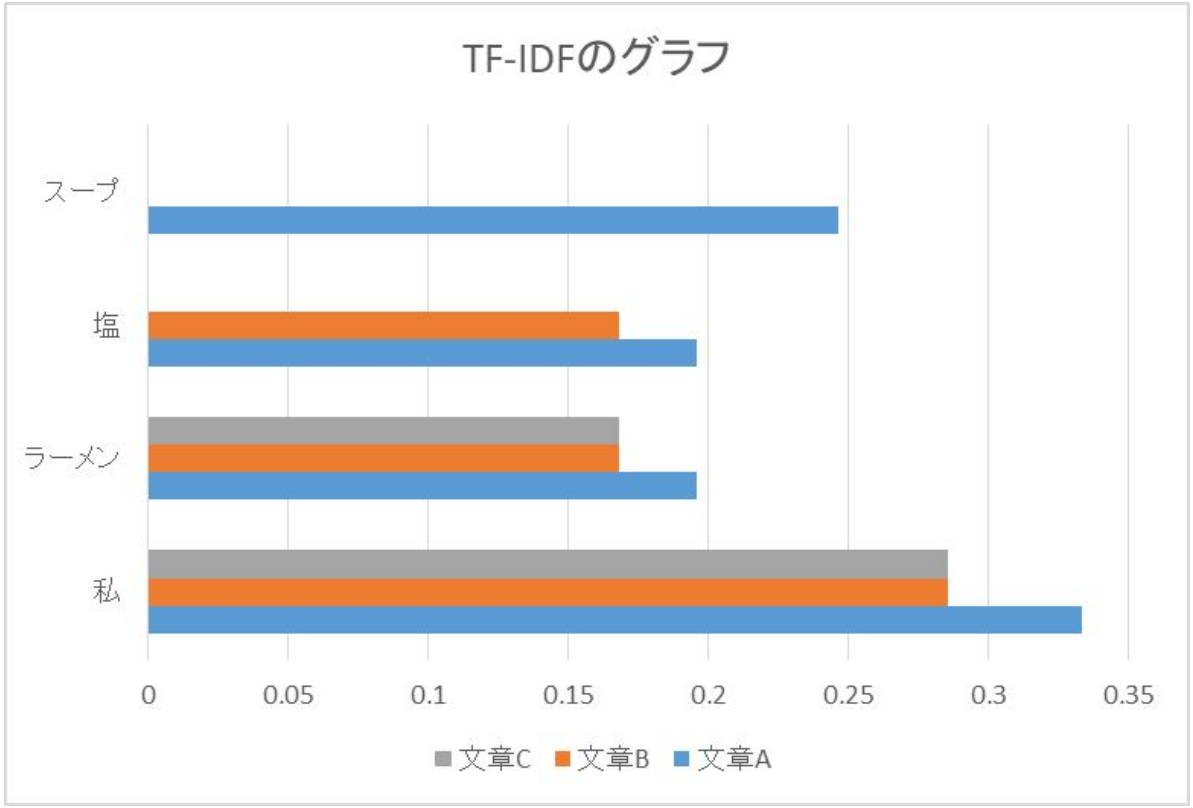
IDF

	私	ラーメン	塩	スープ
文章A	1	1.1761	1.1761	1.4771

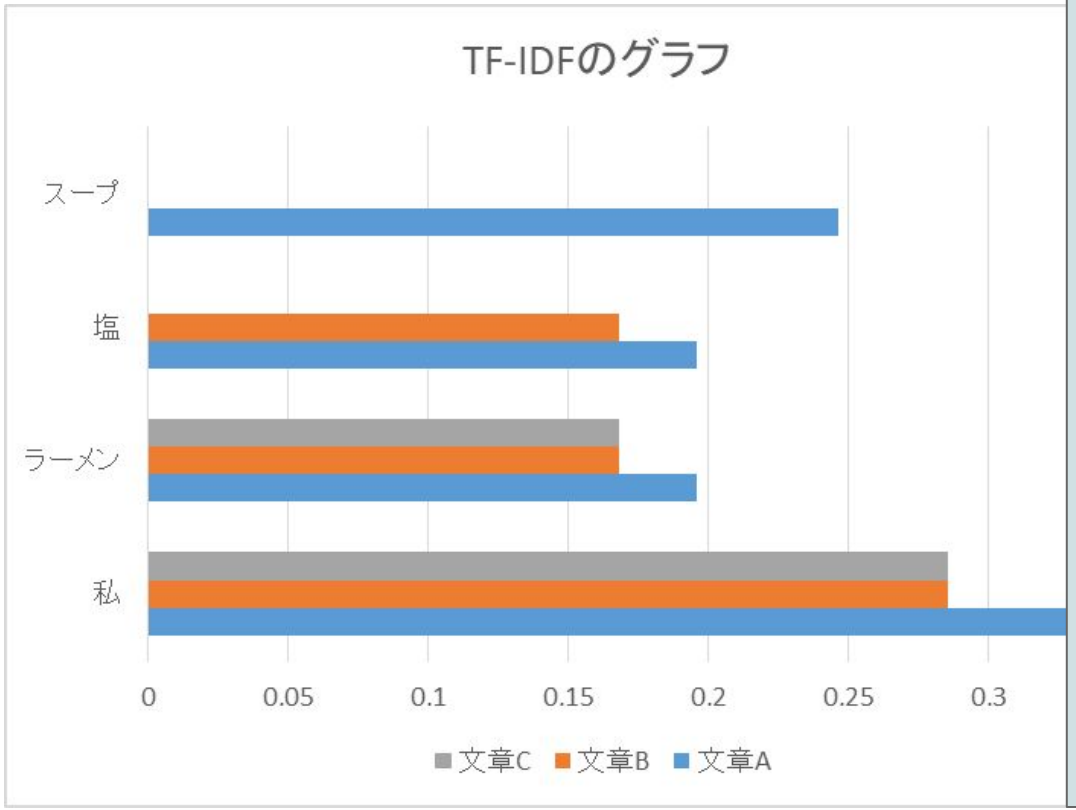
TF-IDF

	私	ラーメン	塩	スープ
文章A	0.3333	0.196	0.196	0.2462
文章B	0.2857	0.168	0.168	0
文章C	0.2857	0.168	0	0

結果をグラフにしてみると……



結果をグラフにしてみると……



- 「私」「ラーメン」はどの文章でもTF-IDFの値が大きい
  - 全体のテーマ
- 「塩」は文章A、Bにて登場
  - この文章は味について述べている?
- 「スープ」は文章Aにのみ登場
  - 文章Aが味、文章Bでは「塩ラーメン」が重要単語だったかも
- 文章Cは味以外の話題で形成されている?

TF-IDFのまとめ

- TFを計算することで、ある文章における重要(そうな)単語を分析できる
- IDFを計算することで、複数の文章の内から希少な単語を見つけることができる

- TF-IDFを計算することで
  複数の文章から重要な単語や、その文章の意味をよく表現しているような特徴的な単語を探索することが出来るようになる

TF-IDFだけで、なんでもできるの?

- TF-IDFはあくまでも特徴的な単語を見つけたり、単語の重要さを数値で表すための手法
  ➔ 情報処理の分野では、さまざまな計算をするための**前処理**

- TF-IDFの計算結果を活用して、例えば以下のことができるようになったりもする
  - 複数の文章から、似ている内容の文章を検索したり
  - 日本語と、それ以外の言語の文章を比べて単語の使い方を学んだり

  ↑のようなことをするためには、更に別の手法が必要になる

TF-IDFで出来そうなことを考えてみる

- レビュー文からネタバレを排除する
  - 「面白い」「良かった」などのIDF値が低そうな単語を含む文章はそのままに、「犯人の〇〇」「ラストの〇〇が」などの特定の人物(≒ネタバレしたい人たち)だけが沢山発言している、IDF値が高そうな単語を含む文章を検出してみる

- シングル楽曲ごとの特徴語を探して、世代間での流行を調査する
  - 誰もが普遍的に扱うテーマや、特定の歌手が大事にしているフレーズ、昔は流行っていたキーワードなんかが分かるかも