検索技術

最終課題

20X3027　大野秀将

一応GitHubを用いたのですができているか心配なのでここに課題を載せておきます。

下のURLはGitHubのURLであり、そこにはプログラムと用いたテキストとここにある文章と同じものがあります。

https://github.com/disneyono/oono

1. データの準備

青空文庫の技術・工学の電気工学にある4つのデータを用いました。

　そのデータは「ランプのいろいろ」、「ムーア灯」、「無線電信の近状」、「写真電送の新法」の4つのデータを用いました。そして、作者を統一して、タイトルによってどのような違いがあるか確認しました。作者は寺田寅彦さんとしました。

２．分析手法と分析の結果

　演習の夏目漱石、芥川龍之介、太宰治の分析のようにまず名詞の語句、をその語句の出現頻度をカウントし、その後にモデルを作成し、そのモデルを利用して類似語を分析し、出力しました。

1. 分析結果の表示と考察

　このデータは同じ作者の時に書く事柄が異なる時、どのような違いがあるか気になり、類似語を分析することで、どのような違いがあるか確認するためにこのような手法を用いました。

結果

「ランプのいろいろ」の単語の頻出数

の(22) これ(20) 灯(20) 光(16) 中(14) 炭(12) ランプ(11) 事(11) もの(11) 金属(10) よう(9) 瓦斯(9) 電流(9) 水銀(9) 色(9) 棒(8) 線(8) 日(8) 年(7) 炭素(6) ［＃(6) 十(6) 灯用(6) 弧(6) 針金(6) 字(5) 白熱(5) 物(5) 普通(5) 電灯(5) 用(5) 電気(5) 空気(5) 管(5) 一(4) 夜(4) 写真(4) 二(4) ガラス(4) 電力(4) 月(4) 寺田(3) 寅彦(3) 例(3) 入力(3) ９(3) 下げ(3) 今(3) 昼(3) 灯火(3)

　ランプのことについて書いてあるので、灯、光、ランプの単語が多く出ている。

「ムーア灯」の単語の頻出数

灯(12) 光(11) 瓦斯(8) 中(7) 電灯(7) 管(7) 電流(7) 一(6) 年(6) 管内(6) の(5) もの(5) 空気(5) 稀薄(5) 色(5) 点(5) 日(5) 十(4) よう(4) 事(4) これ(4) ため(4) 度(4) 故(4) 弁(4) 月(4) ムーア(3) 寺田(3) 寅彦(3) 入力(3) 者(3) 普通(3) 室(3) 硝子(3) .(3) -------------------------------------------------------(2) 例(2) 酔(2) 興(2) きょう(2) 説明(2) 地(2) １(2) 字(2) 上げ(2) さ(2) 数(2) 尺(2) 二(2) 発明

　ムーア灯のことについて書いてあるので、灯、光、電灯の単語が多く出ている。

「無線電信の近状」の単語の頻出数

もの(16) 波(15) 電波(15) これ(11) の(11) 事(10) 器(10) 受信(8) アンテナ(8) 用(6) 線(6) 電話(6) よう(5) 白金(5) 電流(5) 年(5) 日(5) 無線(4) 電信(4) 傍点(4) 発信(4) 所(4) 電気(4) 火花(4) 時(4) 通信(4) 従来(4) 式(4) 金属(4) 間(4) マルコニ(4) 十(4) 筒(4) 月(4) 寺田(3) 寅彦(3) 中(3) 入力(3) 一つ(3) 波動(3) 板(3) 利用(3) 物(3) 一(3) 二(3) 三(3) 四(3) 符号(3) 電路(3) 接続

　無線電信のことについて書いてあるので、電波、受信、アンテナの単語が多く出ている。

「写真電送の新法」の単語の頻出数

機(12) 写真(9) これ(9) もの(8) の(8) 事(7) 式(7) 円筒(7) 十(7) 膜(6) 年(6) 受信(6) 地(5) コルン(5) よう(5) セレニウム(5) 電流(5) 日(5) 電送(4) 中(4) 一(4) 成(4) 効(4) 筒(4) 廻転(4) 原板(4) 写し(4) センチメートル(4) 月(4) 寺田(3) 寅彦(3) 入力(3) 本(3) 目的(3) 光線(3) 濃淡(3) 発信(3) 板(3) 針(3) ペン(3) 八(3) .(3) -------------------------------------------------------(2) 例(2) フィルム(2) １(2) 字(2) 上げ(2) 針金(2) 発明

　写真電送のことについて書いてあるので、写真の単語が多くなっている。

　上記のように、それぞれのタイトルにちなんだ単語が多く出てきていることが分かる。

類似語を分析した結果

「ランプのいろいろ」に出てくる電流の類似語

電流 = 、,中,。,通じる,弧,する,始める,通る,強い,ある

「ムーア灯」に出てくる電流の類似語

電流 = 、,。,なる,管内,光,瓦斯,仕掛け,灯,管,弁

「無線電信の近状」に出てくる電流の類似語

電流 = 線,交番,、,電波,。,これ,する,電話,もの,器

「写真電送の新法」に出てくる電流の類似語

電流 = 、,膜,。,これ,地,する,発信,送る,受信,つける

　上記の結果を見ると、ランプのいろいろのところでは動詞が多くランプの中に電流が通る、強いなどの電気に関することが類似語になっている。ムーア灯では、管内、光、灯などムーア灯に関する言葉が出てきている。そして、無線電信の近状では、線、電波など無線電信に関する単語が多くなっている。写真電送の新法では、発信、受信、送るなど写真電送に関する単語が多く出てきている。このように同じ作者でもタイトルが異なるだけで、頻出単語や電流の類似語が異なってくる。

考察

　しっかりとそれぞれのタイトル通りの単語が頻出単語として出てきており、電流の類似語もそれぞれのタイトル通りの類似語が出てきている。これは、作者が同じでもタイトルが異なれば、頻出単語と類似語は異なってきており、タイトルを確認すればその文章がどのようなことを書いているかが分かると言える。

参考文献

　検索技術のpdfとプログラム

　青空文庫　https://www.aozora.gr.jp/　2022年7月16日　10時15分