情報システム論実習0630課題

課題１

※作成したプログラムは添付した0630.ipynbを参照

【利用した表現手法】

集合ベース手法(Jaccard係数,Sørensen-Dice係数)、ベクトルベース手法（BoW）

【利用した距離尺度】

ユークリッド距離、コサイン類似度

【入力ドキュメントとその実行結果】

入力ドキュメントは以下の通り、１０匹の動物について英英辞典に書かれていたものを使用した。コメントでは文の横に何の動物についての文章か記入してある。

スクリーンショットの画面

自動的に生成された説明

実行結果は以下の通り

集合ベース（Jaccard係数）

|  |
| --- |
| jaccard(doc 0 ,doc 1 ) = 0.2727272727272727  jaccard(doc 0 ,doc 2 ) = 0.07142857142857142  jaccard(doc 0 ,doc 3 ) = 0.16666666666666666  jaccard(doc 0 ,doc 4 ) = 0.13333333333333333  jaccard(doc 0 ,doc 5 ) = 0.13333333333333333  jaccard(doc 0 ,doc 6 ) = 0.05263157894736842  jaccard(doc 0 ,doc 7 ) = 0.045454545454545456  jaccard(doc 0 ,doc 8 ) = 0.14285714285714285  jaccard(doc 0 ,doc 9 ) = 0.07692307692307693  jaccard(doc 1 ,doc 2 ) = 0.18181818181818182  jaccard(doc 1 ,doc 3 ) = 0.1875  jaccard(doc 1 ,doc 4 ) = 0.15384615384615385  jaccard(doc 1 ,doc 5 ) = 0.07142857142857142  jaccard(doc 1 ,doc 6 ) = 0.058823529411764705  jaccard(doc 1 ,doc 7 ) = 0.05  jaccard(doc 1 ,doc 8 ) = 0.07692307692307693  jaccard(doc 1 ,doc 9 ) = 0.09090909090909091  jaccard(doc 2 ,doc 3 ) = 0.1111111111111111  jaccard(doc 2 ,doc 4 ) = 0.14285714285714285  jaccard(doc 2 ,doc 5 ) = 0.14285714285714285  jaccard(doc 2 ,doc 6 ) = 0.11764705882352941  jaccard(doc 2 ,doc 7 ) = 0.047619047619047616  jaccard(doc 2 ,doc 8 ) = 0.07142857142857142  jaccard(doc 2 ,doc 9 ) = 0.0  jaccard(doc 3 ,doc 4 ) = 0.047619047619047616  jaccard(doc 3 ,doc 5 ) = 0.2222222222222222  jaccard(doc 3 ,doc 6 ) = 0.13636363636363635  jaccard(doc 3 ,doc 7 ) = 0.12  jaccard(doc 3 ,doc 8 ) = 0.05  jaccard(doc 3 ,doc 9 ) = 0.0  jaccard(doc 4 ,doc 5 ) = 0.058823529411764705  jaccard(doc 4 ,doc 6 ) = 0.10526315789473684  jaccard(doc 4 ,doc 7 ) = 0.09090909090909091  jaccard(doc 4 ,doc 8 ) = 0.13333333333333333  jaccard(doc 4 ,doc 9 ) = 0.0  jaccard(doc 5 ,doc 6 ) = 0.10526315789473684  jaccard(doc 5 ,doc 7 ) = 0.09090909090909091  jaccard(doc 5 ,doc 8 ) = 0.0625  jaccard(doc 5 ,doc 9 ) = 0.0  jaccard(doc 6 ,doc 7 ) = 0.08  jaccard(doc 6 ,doc 8 ) = 0.1111111111111111  jaccard(doc 6 ,doc 9 ) = 0.0  jaccard(doc 7 ,doc 8 ) = 0.045454545454545456  jaccard(doc 7 ,doc 9 ) = 0.0  jaccard(doc 8 ,doc 9 ) = 0.0 |

集合ベース手法（Sørensen-Dice係数）

|  |
| --- |
| dice(doc 0 ,doc 1 ) = 0.42857142857142855  dice(doc 0 ,doc 2 ) = 0.13333333333333333  dice(doc 0 ,doc 3 ) = 0.2857142857142857  dice(doc 0 ,doc 4 ) = 0.23529411764705882  dice(doc 0 ,doc 5 ) = 0.23529411764705882  dice(doc 0 ,doc 6 ) = 0.1  dice(doc 0 ,doc 7 ) = 0.08695652173913043  dice(doc 0 ,doc 8 ) = 0.25  dice(doc 0 ,doc 9 ) = 0.14285714285714285  dice(doc 1 ,doc 2 ) = 0.3076923076923077  dice(doc 1 ,doc 3 ) = 0.3157894736842105  dice(doc 1 ,doc 4 ) = 0.26666666666666666  dice(doc 1 ,doc 5 ) = 0.13333333333333333  dice(doc 1 ,doc 6 ) = 0.1111111111111111  dice(doc 1 ,doc 7 ) = 0.09523809523809523  dice(doc 1 ,doc 8 ) = 0.14285714285714285  dice(doc 1 ,doc 9 ) = 0.16666666666666666  dice(doc 2 ,doc 3 ) = 0.2  dice(doc 2 ,doc 4 ) = 0.25  dice(doc 2 ,doc 5 ) = 0.25  dice(doc 2 ,doc 6 ) = 0.21052631578947367  dice(doc 2 ,doc 7 ) = 0.09090909090909091  dice(doc 2 ,doc 8 ) = 0.13333333333333333  dice(doc 2 ,doc 9 ) = 0.0  dice(doc 3 ,doc 4 ) = 0.09090909090909091  dice(doc 3 ,doc 5 ) = 0.36363636363636365  dice(doc 3 ,doc 6 ) = 0.24  dice(doc 3 ,doc 7 ) = 0.21428571428571427  dice(doc 3 ,doc 8 ) = 0.09523809523809523  dice(doc 3 ,doc 9 ) = 0.0  dice(doc 4 ,doc 5 ) = 0.1111111111111111  dice(doc 4 ,doc 6 ) = 0.19047619047619047  dice(doc 4 ,doc 7 ) = 0.16666666666666666  dice(doc 4 ,doc 8 ) = 0.23529411764705882  dice(doc 4 ,doc 9 ) = 0.0  dice(doc 5 ,doc 6 ) = 0.19047619047619047  dice(doc 5 ,doc 7 ) = 0.16666666666666666  dice(doc 5 ,doc 8 ) = 0.11764705882352941  dice(doc 5 ,doc 9 ) = 0.0  dice(doc 6 ,doc 7 ) = 0.14814814814814814  dice(doc 6 ,doc 8 ) = 0.2  dice(doc 6 ,doc 9 ) = 0.0  dice(doc 7 ,doc 8 ) = 0.08695652173913043  dice(doc 7 ,doc 9 ) = 0.0  dice(doc 8 ,doc 9 ) = 0.0 |

ベクトルベース手法（ユークリッド距離）

|  |
| --- |
| euclidean\_distance(bow 0 ,bow 1 ) = 2.8284271247461903  euclidean\_distance(bow 0 ,bow 2 ) = 3.605551275463989  euclidean\_distance(bow 0 ,bow 3 ) = 3.872983346207417  euclidean\_distance(bow 0 ,bow 4 ) = 3.605551275463989  euclidean\_distance(bow 0 ,bow 5 ) = 3.605551275463989  euclidean\_distance(bow 0 ,bow 6 ) = 4.242640687119285  euclidean\_distance(bow 0 ,bow 7 ) = 4.58257569495584  euclidean\_distance(bow 0 ,bow 8 ) = 3.4641016151377544  euclidean\_distance(bow 0 ,bow 9 ) = 3.4641016151377544  euclidean\_distance(bow 1 ,bow 2 ) = 3.0  euclidean\_distance(bow 1 ,bow 3 ) = 3.605551275463989  euclidean\_distance(bow 1 ,bow 4 ) = 3.3166247903554  euclidean\_distance(bow 1 ,bow 5 ) = 3.605551275463989  euclidean\_distance(bow 1 ,bow 6 ) = 4.0  euclidean\_distance(bow 1 ,bow 7 ) = 4.358898943540674  euclidean\_distance(bow 1 ,bow 8 ) = 3.4641016151377544  euclidean\_distance(bow 1 ,bow 9 ) = 3.1622776601683795  euclidean\_distance(bow 2 ,bow 3 ) = 4.0  euclidean\_distance(bow 2 ,bow 4 ) = 3.4641016151377544  euclidean\_distance(bow 2 ,bow 5 ) = 3.4641016151377544  euclidean\_distance(bow 2 ,bow 6 ) = 3.872983346207417  euclidean\_distance(bow 2 ,bow 7 ) = 4.47213595499958  euclidean\_distance(bow 2 ,bow 8 ) = 3.605551275463989  euclidean\_distance(bow 2 ,bow 9 ) = 3.605551275463989  euclidean\_distance(bow 3 ,bow 4 ) = 4.47213595499958  euclidean\_distance(bow 3 ,bow 5 ) = 3.7416573867739413  euclidean\_distance(bow 3 ,bow 6 ) = 4.358898943540674  euclidean\_distance(bow 3 ,bow 7 ) = 4.69041575982343  euclidean\_distance(bow 3 ,bow 8 ) = 4.358898943540674  euclidean\_distance(bow 3 ,bow 9 ) = 4.358898943540674  euclidean\_distance(bow 4 ,bow 5 ) = 4.0  euclidean\_distance(bow 4 ,bow 6 ) = 4.123105625617661  euclidean\_distance(bow 4 ,bow 7 ) = 4.47213595499958  euclidean\_distance(bow 4 ,bow 8 ) = 3.605551275463989  euclidean\_distance(bow 4 ,bow 9 ) = 3.872983346207417  euclidean\_distance(bow 5 ,bow 6 ) = 4.123105625617661  euclidean\_distance(bow 5 ,bow 7 ) = 4.47213595499958  euclidean\_distance(bow 5 ,bow 8 ) = 3.872983346207417  euclidean\_distance(bow 5 ,bow 9 ) = 3.872983346207417  euclidean\_distance(bow 6 ,bow 7 ) = 4.795831523312719  euclidean\_distance(bow 6 ,bow 8 ) = 4.0  euclidean\_distance(bow 6 ,bow 9 ) = 4.242640687119285  euclidean\_distance(bow 7 ,bow 8 ) = 4.58257569495584  euclidean\_distance(bow 7 ,bow 9 ) = 4.58257569495584  euclidean\_distance(bow 8 ,bow 9 ) = 3.7416573867739413 |

ベクトルベース手法（コサイン類似度）

|  |
| --- |
| cosine\_similarity(bow 0 ,bow 1 ) = 0.43301270189221935  cosine\_similarity(bow 0 ,bow 2 ) = 0.13363062095621217  cosine\_similarity(bow 0 ,bow 3 ) = 0.294174202707276  cosine\_similarity(bow 0 ,bow 4 ) = 0.2357022603955158  cosine\_similarity(bow 0 ,bow 5 ) = 0.2357022603955158  cosine\_similarity(bow 0 ,bow 6 ) = 0.10206207261596574  cosine\_similarity(bow 0 ,bow 7 ) = 0.09128709291752767  cosine\_similarity(bow 0 ,bow 8 ) = 0.24999999999999994  cosine\_similarity(bow 0 ,bow 9 ) = 0.14433756729740646  cosine\_similarity(bow 1 ,bow 2 ) = 0.3086066999241838  cosine\_similarity(bow 1 ,bow 3 ) = 0.33968311024337877  cosine\_similarity(bow 1 ,bow 4 ) = 0.2721655269759087  cosine\_similarity(bow 1 ,bow 5 ) = 0.13608276348795434  cosine\_similarity(bow 1 ,bow 6 ) = 0.11785113019775793  cosine\_similarity(bow 1 ,bow 7 ) = 0.10540925533894598  cosine\_similarity(bow 1 ,bow 8 ) = 0.14433756729740646  cosine\_similarity(bow 1 ,bow 9 ) = 0.16666666666666669  cosine\_similarity(bow 2 ,bow 3 ) = 0.20965696734438366  cosine\_similarity(bow 2 ,bow 4 ) = 0.2519763153394848  cosine\_similarity(bow 2 ,bow 5 ) = 0.2519763153394848  cosine\_similarity(bow 2 ,bow 6 ) = 0.2182178902359924  cosine\_similarity(bow 2 ,bow 7 ) = 0.0975900072948533  cosine\_similarity(bow 2 ,bow 8 ) = 0.13363062095621217  cosine\_similarity(bow 2 ,bow 9 ) = 0.0  cosine\_similarity(bow 3 ,bow 4 ) = 0.09245003270420486  cosine\_similarity(bow 3 ,bow 5 ) = 0.36980013081681945  cosine\_similarity(bow 3 ,bow 6 ) = 0.24019223070763074  cosine\_similarity(bow 3 ,bow 7 ) = 0.21483446221182986  cosine\_similarity(bow 3 ,bow 8 ) = 0.09805806756909201  cosine\_similarity(bow 3 ,bow 9 ) = 0.0  cosine\_similarity(bow 4 ,bow 5 ) = 0.1111111111111111  cosine\_similarity(bow 4 ,bow 6 ) = 0.19245008972987526  cosine\_similarity(bow 4 ,bow 7 ) = 0.17213259316477406  cosine\_similarity(bow 4 ,bow 8 ) = 0.2357022603955158  cosine\_similarity(bow 4 ,bow 9 ) = 0.0  cosine\_similarity(bow 5 ,bow 6 ) = 0.19245008972987526  cosine\_similarity(bow 5 ,bow 7 ) = 0.17213259316477406  cosine\_similarity(bow 5 ,bow 8 ) = 0.1178511301977579  cosine\_similarity(bow 5 ,bow 9 ) = 0.0  cosine\_similarity(bow 6 ,bow 7 ) = 0.14907119849998599  cosine\_similarity(bow 6 ,bow 8 ) = 0.20412414523193148  cosine\_similarity(bow 6 ,bow 9 ) = 0.0  cosine\_similarity(bow 7 ,bow 8 ) = 0.09128709291752767  cosine\_similarity(bow 7 ,bow 9 ) = 0.0  cosine\_similarity(bow 8 ,bow 9 ) = 0.0 |

【実行結果の考察】

Sørensen-Dice係数の結果をみると数値が高いものとして、

（doc0, doc1）=0.42857142857142855

(doc1, doc3) = 0.3157894736842105

などが挙げられるが、doc0は猫、doc1は犬、doc3は象であり、いずれも説明の中に四本足であることが書かれている。そのため結果の数値が高かったのだと考えられる。四本足について書かれていなくても、哺乳類であれば0.1〜0.2ほどの結果が出ているため、ある程度、種族によって書かれる内容に関係性があるのではないかと考えられる。

逆に数値が0のものも存在する。

(doc 2 ,doc 9 ) = 0.0

(doc 8 ,doc 9 ) = 0.0

Doc2は魚、doc8は蝙蝠、doc9は鶏

どれも種族が違うことがわかる。Doc9は他の値も０のものがいくつか見受けられるが、今回洗濯した動物の中に鳥類は鶏しかいないことからそれが原因なのではないかと考えられる。

これらの傾向は他の実行結果でも同様にみることができる。

このことから、動物の説明から種族の関係性をある程度読み解くことができているのではないかと考察した。

【感想】

今回初めて、テキストマイニングを自分で実装することができ、今まで知識としてしかわかっていなかったものを実際に使ってみるという体験をすることができた。テキストマイニングをするにあたって、どこまで前処理すべきなのか、どのような手法を選択すべきなのかと言った考えることがたくさんあった。今回はあくまで触りのようなことしかやっていないが、テキストマイニングの可能性は感じることができたため、今後機会があれば、自身で学習して、より高度な内容が扱えたら良いなと感じた。