**Clustering**

1. **Toy dataset Bag of Words:**

K-means algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная | Жаккара |
| Все части речи | Индекс Rand:  0,45 0,46 0,56  Индекс Adjusted Rand:  -0,08 -0,07 0,13  Однородность:  0,06 0,07 0,15  Полнота:  0,14 0,15 0,33 | Индекс Rand:  0,45 0,6 1  Индекс Adjusted Rand:  -0,09 0,21 1  Однородность:  0,001 0,3 1  Полнота:  0,001 0,3 1 | Индекс Rand:  0 0,009 0,49  Индекс Adjusted Rand:  -0,02 0,0004 0  Однородность:  0 0,002 0,08  Полнота:  0 0,001 0,07 |
| Существительные и прилагательные | Индекс Rand:  0.47 0.55 0.8  Индекс Adjusted Rand:  -0.05 0.13 0.6  Однородность:  0.008 0.17 0.57  Полнота:  0.01 0.32 0.63 | Индекс Rand:  0.44 0.58 0.8  Индекс Adjusted Rand:  -0.12 0.17 0.6  Однородность:  0 0.25 0.63  Полнота:  0 0.29 0.63 | Индекс Rand:  0 0.01 0.64  Индекс Adjusted Rand:  0 0.005 0.29  Однородность:  0 0.009 0.43  Полнота:  0 0.009 0.43 |
| Существительные | Индекс Rand:  0.53  Индекс Adjusted Rand:  0.1  Однородность:  0.15  Полнота:  0.31 | Индекс Rand:  0.44 0.59 0.8  Индекс Adjusted Rand:  -0.11 0.19 0.6  Однородность:  0.006 0.27 0.57  Полнота:  0.007 0.31 0.63 | Индекс Rand:  0 0.009 0.47  Индекс Adjusted Rand:  -0.05 -0.0009 0  Однородность:  0 0.0002 0.008  Полнота:  0 0.0002 0.01 |

Hierarchical algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная | Жаккара |
| Все части речи | Индекс Rand:  0.45  Индекс Adjusted Rand:  -0.08  Однородность:  0.07  Полнота:  0.14 | Индекс Rand:  0,81  Индекс Adjusted Rand:  0,64  Однородность:  0,58  Полнота:  0,65 | Индекс Rand:  0,45  Индекс Adjusted Rand:  -0,09  Однородность:  0,004  Полнота:  0,003 |
| Существительные и прилагательные | Индекс Rand:  0.53  Индекс Adjusted Rand:  0.1  Однородность:  0.15  Полнота:  0.31 | Индекс Rand:  0.8  Индекс Adjusted Rand:  0.6  Однородность:  0.57  Полнота:  0.63 | Индекс Rand:  0.44  Индекс Adjusted Rand:  -0.11  Однородность:  0.006  Полнота:  0.007 |
| Существительные | Индекс Rand:  0.53  Индекс Adjusted Rand:  0.1  Однородность:  0.15  Полнота:  0.31 | Индекс Rand:  0.8  Индекс Adjusted Rand:  0.6  Однородность:  0.57  Полнота:  0.63 | Индекс Rand:  0.47  Индекс Adjusted Rand:  -0.07  Однородность:  0.05  Полнота:  0.05 |

Вывод: в данном эксперименте наиболее успешным стал алгоритм k-means кластеризации с косинусной метрикой на всех частях речи, так как он получил лучшие значения в выбранных метриках. Наиболее подходящей метрикой во всех случаях является косинусная. Если брать средние показатели, то выигрывает алгоритм иерархической кластеризации с косинусной метрикой на всех частях речи

1. **20NewsGroups TfIdf:**

**Хорошо различимые классы**

K-means algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0.43 0.59 0.65  V measure:  0.21 0.34 0.41  Mutual info score:  0.2 0.34 0.41  Fowlkes Mallows:  0.46 0.5 0.53 | Adjuster Rand:  0.58 0.73 0.79  V measure:  0.27 0.43 0.55  Mutual info score:  0.27 0.43 0.55  Fowlkes Mallows:  0.47 0.59 0.65 |
| Существительные и прилагательные | Adjuster Rand:  0.45 0.62 0.72  V measure:  0.24 0.38 0.46  Mutual info score:  0.24 0.38 0.46  Fowlkes Mallows:  0.45 0.51 0.59 | Adjuster Rand:  0.63 0.77 0.79  V measure:  0.22 0.47 0.54  Mutual info score:  0.22 0.47 0.54  Fowlkes Mallows:  0.45 0.62 0.65 |
| Существительные | Adjuster Rand:  0.48 0.59 0.65  V measure:  0.22 0.35 0.42  Mutual info score:  0.22 0.35 0.42  Fowlkes Mallows:  0.45 0.5 0.53 | Adjuster Rand:  0.72 0.77 0.82  V measure:  0.43 0.5 0.55  Mutual info score:  0.42 0.5 0.55  Fowlkes Mallows:  0.55 0.63 0.65 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.51 c:0.38 a:0.25 s:0.25  V measure:  w:0.14 c:0.01 a:0 s:0  Mutual info score:  w:0.23 c:0.06 a:0.0008 s:0.0008  Fowlkes Mallows:  w:0.48 c:0.15 a:0.15 s:0.15 | Adjuster Rand:  c:0.26 a:0.25 s:0.25  V measure:  c:0 a:0 s:0  Mutual info score:  c:0.004 a:0.0008 s:0.0008  Fowlkes Mallows:  c:0.1 a:0.15 s:0.15 |
| Существительные и прилагательные | Adjuster Rand:  w:0.49 c:0.48 a:0.25 s:0.25  V measure:  w:0.12 c:0.03 a:0 s:0  Mutual info score:  w:0.22 c:0.02 a:0.008 s:0.008  Fowlkes Mallows:  w:0.47 c:0.09 a:0.15 s:0.15 | Adjuster Rand:  c:0.26 a:0.25 s:0.25  V measure:  c:0 a:0 s:0  Mutual info score:  c:0.009 a:0.0008 s:0.0008  Fowlkes Mallows:  c:.0.2 a:0.15 s:0.15 |
| Существительные | Adjuster Rand:  w:0.4 c:0.34 a:0.28 s:0.28  V measure:  w:0.04 c:0.005 a:0.00007 s:0.00007  Mutual info score:  w:0.12 c:0.05 a:0.003 s:0.003  Fowlkes Mallows:  w:0.3 c:0.16 a:0.02 s:0.02 | Adjuster Rand:  c:0.26 a:0.28 s:0.28  V measure:  c:0.00002 a:0.00007 s:0.00007  Mutual info score:  c:0.004 a:0.003 s:0.003  Fowlkes Mallows:  c:0.1 a:0.02 s:0.02 |

**Плохо различимые классы**

K-means algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0,47 0,57 0,63  V measure:  0,19 0,25 0,30  Mutual info score:  0,19 0,25 0,30  Fowlkes Mallows:  0,41 0,45 0,51 | Adjuster Rand:  0,56 0,65 0,73  V measure:  0,13 0,28 0,38  Mutual info score:  0,13 0,28 0,38  Fowlkes Mallows:  0,38 0,48 0,56 |
| Существительные и прилагательные | Adjuster Rand:  0.49 0.55 0.60  V measure:  0.22 0.27 0.32  Mutual info score:  0.22 0.27 0.32  Fowlkes Mallows:  0.42 0.47 0.52 | Adjuster Rand:  0.61 0.68 0.74  V measure:  0.23 0.31 0.39  Mutual info score:  0.23 0.31 0.39  Fowlkes Mallows:  0.44 0.50 0.57 |
| Существительные | Adjuster Rand:  0,47 0,56 0,62  V measure:  0,08 0,26 0,32  Mutual info score:  0,08 0,26 0,32  Fowlkes Mallows:  0,40 0,46 0,53 | Adjuster Rand:  0,61 0,67 0,71  V measure:  0,19 0,28 0,37  Mutual info score:  0,19 0,28 0,37  Fowlkes Mallows:  0,41 0,48 0,57 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.48 c:0.58 a:0.34 s:0.34  V measure:  w:0.03 c:0.03 a:0. s:0.  Mutual info score:  w:0.16 c:0.05 a:0. s:0.  Fowlkes Mallows:  w:0.22 c:0.04 a:0.12 s:0.13 | Adjuster Rand:  c:0.34 a:0.34 s:0.34  V measure:  c:0. a:0. s:0.  Mutual info score:  c:0. a:0. s:0.  Fowlkes Mallows:  c:0.06 a:0.12 s:0.13 |
| Существительные и прилагательные | Adjuster Rand:  w:0.49 c:0.54 a:0.34 s:0.34  V measure:  w:0.05 c:0.01 a:0. s:0.  Mutual info score:  w:0.18 c:0.01 a:0. s:0.  Fowlkes Mallows:  w:0.25 c:0.01 a:0.13 s:0.13 | Adjuster Rand:  c:0.34 a:0.34 s:0.34  V measure:  c:0. a:0. s:0.  Mutual info score:  c:0. a:0. s:0.  Fowlkes Mallows:  c:0.07 a:0.13 s:0.13 |
| Существительные | Adjuster Rand:  w:0.51 c:0.42 a:0.34 s:0.34  V measure:  w:0.05 c:0.01 a:0. s:0.  Mutual info score:  w:0.19 c:0.02 a:0. s:0.  Fowlkes Mallows:  w:0.23 c:0.06 a:0.13 s:0.13 | Adjuster Rand:  c:0.34 a:0.34 s:0.34  V measure:  c:0. a:0. s:0.  Mutual info score:  c:0. a:0. s:0.  Fowlkes Mallows:  c:0.05 a:0.13 s:0.13 |

Вывод: на полноценном наборе данных с хорошо различимыми классами наиболее успешным оказался алгоритм k-means кластеризации с косинусной метрикой на существительных. На плохо различимых классах лучше всего себя показал также алгоритм k-means кластеризации с косинусной метрикой, но на существительных и прилагательных.

1. **20NewGroups Bert:**

**Хорошо различимые данные**

**Base Uncased**

K-means algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0.61 0.61 0.62  V measure:  0.05 0.05 0.16  Mutual info score:  0.05 0.05 0.16  Fowlkes Mallows:  0.31 0.32 0.38 | Adjuster Rand:  0.53 0.60 0.61  V measure:  0.03 0.05 0.05  Mutual info score:  0.03 0.04 0.05  Fowlkes Mallows:  0.31 0.32 0.36 |
| Существительные и прилагательные | Adjuster Rand:  0.63 0.63 0.63  V measure:  0.19 0.19 0.19  Mutual info score:  0.19 0.19 0.19  Fowlkes Mallows:  0.40 0.40 .40 | Adjuster Rand:  0.63 0.62 0.63  V measure:  0.03 0.19 0.19  Mutual info score:  0.03 0.18 0.19  Fowlkes Mallows:  0.36 0.40 0.40 |
| Существительные | Adjuster Rand:  0,62 0,62 0,62  V measure:  0,18 0,18 0,19  Mutual info score:  0,18 0,18 0,19  Fowlkes Mallows:  0,39 0,39 0,39 | Adjuster Rand:  0,62 0,62 0,62  V measure:  0,19 0,19 0,19  Mutual info score:  0,18 0,19 0,19  Fowlkes Mallows:  0,39 0,39 0,39 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.58 c:0.44 a:0.28 s:0.28  V measure:  w:0.09 c:0.01 a:0 s:0  Mutual info score:  w:0.12 c:0.03 a:0 s:0  Fowlkes Mallows:  w:0.16 c:0.06 a:0.02 s:0.02 | Adjuster Rand:  c:0.56 a:0.28 s:0.28  V measure:  c:0.02 a:0 s:0  Mutual info score:  c:0.03 a:0 s:0  Fowlkes Mallows:  c:0.04 a:0.02 s:0.02 |
| Существительные и прилагательные | Adjuster Rand:  w:0.64 c:0.40 a:0.28 s:0.28  V measure:  w:0.18 c:0.02 a:0. s:0.  Mutual info score:  w:0.21 c:0.05 a:0. s:0.  Fowlkes Mallows:  w:0.26 c:0.13 a:0.02 s:0.02 | Adjuster Rand:  c:0.48 a:0.28 s:0.28  V measure:  c:0.03 a:0. s:0.  Mutual info score:  c:0.05 a:0. s:0.  Fowlkes Mallows:  c:.0.09 a:0.02 s:0.02 |
| Существительные | Adjuster Rand:  w:0.64 c:0.52 a:0.28 s:0.28  V measure:  w:0.18 c:0.03 a:0. s:0.  Mutual info score:  w:0.23 c:0.03 a:0. s:0.  Fowlkes Mallows:  w:0.28 c:0.05 a:0.02 s:0.02 | Adjuster Rand:  c:0.60 a:0.29 s:0.28  V measure:  c:0.04 a:0. s:0.  Mutual info score:  c:0.04 a:0. s:0.  Fowlkes Mallows:  c:.0.05 a:0.02 s:0.02 |

**Base Cased**

K-means algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0.60 0.60 0.64  V measure:  0.02 0.03 0.05  Mutual info score:  0.02 0.03 0.05  Fowlkes Mallows:  0.27 0.29 0.30 | Adjuster Rand:  0.50 0.59 0.60  V measure:  0.02 0.03 0.03  Mutual info score:  0.02 0.02 0.03  Fowlkes Mallows:  0.29 0.30 0.38 |
| Существительные и прилагательные | Adjuster Rand:  0,58 0,60 0,64  V measure:  0,03 0,04 0,07  Mutual info score:  0,03 0,04 0,06  Fowlkes Mallows:  0,28 0,31 0,31 | Adjuster Rand:  0,59 0,61 0,64  V measure:  0,03 0,04 0,07  Mutual info score:  0,03 0,04 0,07  Fowlkes Mallows:  0,28 0,31 0,31 |
| Существительные | Adjuster Rand:  0.58 0.59 0.64  V measure:  0.03 0.03 0.06  Mutual info score:  0.03 0.03 0.06  Fowlkes Mallows:  0.28 0.31 0.31 | Adjuster Rand:  0.58 0.60 0.64  V measure:  0.03 0.04 0.08  Mutual info score:  0.03 0.04 0.08  Fowlkes Mallows:  0.28 0.31 0.31 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.59 c:0.50 a:0.25 s:0.28  V measure:  w:0.02 c:0.01 a:0 s:0  Mutual info score:  w:0.03 c:0.01 a:0. s:0.  Fowlkes Mallows:  w:0.03 c:0.03 a:0.16 s:0.02 | Adjuster Rand:  c:0.45 a:0.28 s:0.28  V measure:  c:0.01 a:0. s:0.  Mutual info score:  c:0.01 a:0. s:0.  Fowlkes Mallows:  c:.0.03 a:0.03 s:0.03 |
| Существительные и прилагательные | Adjuster Rand:  w:0.58 c:0.39 a:0.25 s:0.25  V measure:  w:0.03 c:0.01 a:0. s:0.  Mutual info score:  w:0.03 c:0.02 a:0. s:0.  Fowlkes Mallows:  w:0.03 c:0.06 a:0.08 s:0.15 | Adjuster Rand:  c:0.50 a:0.25 s:0.28  V measure:  c:0.02 a:0. s:0.  Mutual info score:  c:0.03 a:0. s:0.  Fowlkes Mallows:  c:.0.05 a:0.12 s:0.02 |
| Существительные | Adjuster Rand:  w:0.55 c:0.57 a:0.25 s:0.25  V measure:  w:0.01 c:0.02 a:0. s:0.  Mutual info score:  w:0.02 c:0.03 a:0. s:0.  Fowlkes Mallows:  w:0.03 c:0.04 a:0.08 s:0.15 | Adjuster Rand:  c:0.48 a:0.25 s:0.28  V measure:  c:0. a:0. s:0.  Mutual info score:  c:0.01 a:0. s:0.  Fowlkes Mallows:  c:.0.03 a:0.15 s:0.02 |

**Distile Bert**

K-means algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0.61 0.62 0.62  V measure:  0.06 0.06 0.09  Mutual info score:  0.06 0.06 0.09  Fowlkes Mallows:  0.32 0.32 .07 | Adjuster Rand:  0.63 0.64 0.67  V measure:  0.07 0.07 0.21  Mutual info score:  0.06 0.07 0.20  Fowlkes Mallows:  0.31 0.32 0.39 |
| Существительные и прилагательные | Adjuster Rand:  0,61 0,62 0,66  V measure:  0,06 0,12 0,22  Mutual info score:  0,06 0,12 0,22  Fowlkes Mallows:  0,32 0,35 0,41 | Adjuster Rand:  0,63 0,65 0,67  V measure:  0,08 0,12 0,20  Mutual info score:  0,08 0,12 0,20  Fowlkes Mallows:  0,33 0,34 0,39 |
| Существительные | Adjuster Rand:  0.61 0.62 0.67  V measure:  0.06 0.11 0.20  Mutual info score:  0.06 0.11 0.19  Fowlkes Mallows:  0.32 0.35 0.39 | Adjuster Rand:  0.62 0.66 0.67  V measure:  0.06 0.16 0.20  Mutual info score:  0.06 0.16 0.20  Fowlkes Mallows:  0.31 0.36 0.39 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.60 c:0.57 a:0.39 s:0.25  V measure:  w:0.06 c:0.01 a:0. s:0.  Mutual info score:  w:0.06 c:0.02 a:0.1 s:0.1  Fowlkes Mallows:  w:0.07 c:0.03 a:0.02 s:0.15 | Adjuster Rand:  c:0.59 a:0.43 s:0.25  V measure:  c:0.03 a:0.01 s:0.  Mutual info score:  c:0.06 a:0.02 s:0.  Fowlkes Mallows:  c:.0.07 a:0.05 s:0.15 |
| Существительные и прилагательные | Adjuster Rand:  w:0.64 c:0.60 a:0.43 s:0.25  V measure:  w:0.14 c:0.05 a:0.01 s:0.  Mutual info score:  w:0.13 c:0.05 a:0.01 s:0.  Fowlkes Mallows:  w:0.15 c:0.06 a:0.02 s:0.15 | Adjuster Rand:  c:0.49 a:0.47 s:0.25  V measure:  c:0.01 a:0.01 s:0.  Mutual info score:  c:0.01 a:0.01 s:0.  Fowlkes Mallows:  c:.0.02 a:0.03 s:0.15 |
| Существительные | Adjuster Rand:  w:0.62 c:0.57 a:0.47 s:0.25  V measure:  w:0.11 c:0.07 a:0.01 s:0.  Mutual info score:  w:0.15 c:0.06 a:0.01 s:0.  Fowlkes Mallows:  w:0.18 c:0.09 a:0.03 s:0.15 | Adjuster Rand:  c:0.60 a:0.25 s:0.25  V measure:  c:0.08 a:0. s:0.  Mutual info score:  c:0.08 a:0. s:0.  Fowlkes Mallows:  c:.0.10 a:0.17 s:0.15 |

**Плохо различимые данные**

**Base Uncased**

K-means algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0.57 0.57 0.58  V measure:  0.02 0.02 0.02  Mutual info score:  0.02 0.02 0.02  Fowlkes Mallows:  0.34 0.34 0.34 | Adjuster Rand:  0.57 0.57 0.58  V measure:  0.02 0.02 0.02  Mutual info score:  0.02 0.02 0.02  Fowlkes Mallows:  0.34 0.34 0.34 |
| Существительные и прилагательные | Adjuster Rand:  0.56 0.56 0.56  V measure:  0.03 0.03 0.03  Mutual info score:  0.03 0.03 0.03  Fowlkes Mallows:  0.36 0.36 0.36 | Adjuster Rand:  0.52 0.56 0.56  V measure:  0.02 0.02 0.03  Mutual info score:  0.02 0.02 0.03  Fowlkes Mallows:  0.35 0.35 0.42 |
| Существительные | Adjuster Rand:  0.55 0.56 0.56  V measure:  0.02 0.03 0.03  Mutual info score:  0.02 0.02 0.03  Fowlkes Mallows:  0.36 0.36 0.36 | Adjuster Rand:  0.55 0.55 0.56  V measure:  0.02 0.02 0.02  Mutual info score:  0.02 0.02 0.02  Fowlkes Mallows:  0.36 0.36 0.36 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.55 c:0.50 a:0.35 s:0.35  V measure:  w:0.03 c:0.02 a:0. s:0.  Mutual info score:  w:0.04 c:0.02 a:0. s:0.  Fowlkes Mallows:  w:0.04 c:0.04 a:0. s:0. | Adjuster Rand:  c:0.45 a:0.35 s:0.35  V measure:  c:0. a:0. s:0.  Mutual info score:  c:0.02 a:0. s:0.  Fowlkes Mallows:  c:.0.03 a:0.01 s:0.01 |
| Существительные и прилагательные | Adjuster Rand:  w:0.56 c:0.53 a:0.35 s:0.35  V measure:  w:0.04 c:0.01 a:0. s:0.  Mutual info score:  w:0.03 c:0.02 a:0. s:0.  Fowlkes Mallows:  w:0.03 c:0.02 a:0.01 s:0.01 | Adjuster Rand:  c:0.52 a:0.35 s:0.35  V measure:  c:0.02 a:0. s:0.  Mutual info score:  c:0.02 a:0. s:0.  Fowlkes Mallows:  c:.0.03 a:0.01 s:0.01 |
| Существительные | Adjuster Rand:  w:0.53 c:0.52 a:0.35 s:0.35  V measure:  w:0.01 c:0.03 a:0. s:0.  Mutual info score:  w:0.02 c:0.04 a:0. s:0.  Fowlkes Mallows:  w:0.02 c:0.05 a:0. s:0. | Adjuster Rand:  c:0.52 a:0.35 s:0.35  V measure:  c:0.03 a:0. s:0.  Mutual info score:  c:0.02 a:0. s:0.  Fowlkes Mallows:  c:.0.03 a:0. s:0. |

**Base Cased**

K-means algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0.53 0.57 0.58  V measure:  0.01 0.01 0.02  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.31 0.33 0.38 | Adjuster Rand:  0.53 0.57 0.58  V measure:  0.01 0.01 0.02  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.31 0.34 0.38 |
| Существительные и прилагательные | Adjuster Rand:  0.57 0.57 0.57  V measure:  0.01 0.02 0.02  Mutual info score:  0.02 0.02 0.02  Fowlkes Mallows:  0.33 0.34 0.34 | Adjuster Rand:  0.50 0.57 0.59  V measure:  0.01 0.02 0.02  Mutual info score:  0.01 0.02 0.02  Fowlkes Mallows:  0.30 0.34 0.43 |
| Существительные | Adjuster Rand:  0.49 0.56 0.59  V measure:  0.01 0.01 0.02  Mutual info score:  0.02 0.02 0.02  Fowlkes Mallows:  0.31 0.34 0.44 | Adjuster Rand:  0.49 0.56 0.59  V measure:  0.02 0.02 0.02  Mutual info score:  0.02 0.02 0.02  Fowlkes Mallows:  0.30 0.34 0.44 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.57 c:0.49 a:0.34 s:0.35  V measure:  w:0.01 c:0.01 a:0. s:0.  Mutual info score:  w:0.01 c:0.01 a:0. s:0.  Fowlkes Mallows:  w:0.01 c:0.01 a:0.05 s:0. | Adjuster Rand:  c:0.52 a:0.34 s:0.35  V measure:  c:0.01 a:0. s:0.  Mutual info score:  c:0.01 a:0. s:0.  Fowlkes Mallows:  c:.0.02 a:0.10 s:0. |
| Существительные и прилагательные | Adjuster Rand:  w:0.55 c:0.46 a:0.34 s:0.35  V measure:  w:0.02 c:0.01 a:0. s:0.  Mutual info score:  w:0.02 c:0.01 a:0. s:0.  Fowlkes Mallows:  w:0.02 c:0.03 a:0.13 s:0. | Adjuster Rand:  c:0.45 a:0.34 s:0.34  V measure:  c:0.01 a:0. s:0.  Mutual info score:  c:0.01 a:0. s:0.  Fowlkes Mallows:  c:.0.02 a:0.07 s:0.13 |
| Существительные | Adjuster Rand:  w:0.54 c:0.48 a:0.34 s:0.34  V measure:  w:0.02 c:0.01 a:0. s:0.  Mutual info score:  w:0.02 c:0.01 a:0. s:0.  Fowlkes Mallows:  w:0.02 c:0.02 a:0.14 s:0.14 | Adjuster Rand:  c:0.45 a:0.34 s:0.34  V measure:  c:0.01 a:0. s:0.  Mutual info score:  c:0.01 a:0. s:0.  Fowlkes Mallows:  c:.0.02 a:0.13 s:0.13 |

**Distile Bert**

K-means algorithm

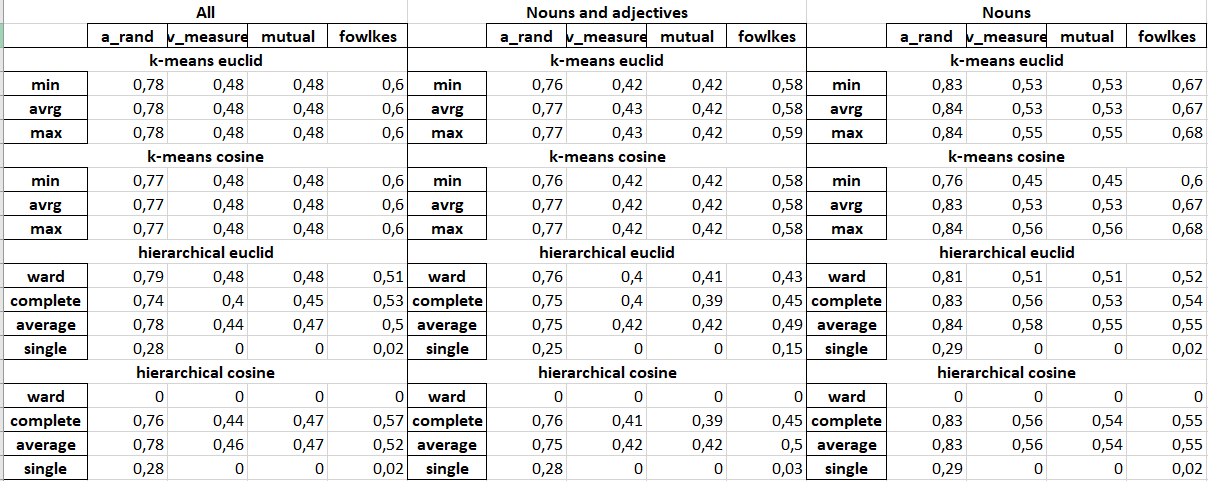
|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  0.58 0.58 0.58  V measure:  0.01 0.01 0.01  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.31 0.31 0.32 | Adjuster Rand:  0.58 0.58 0.58  V measure:  0.01 0.01 0.01  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.31 0.31 0.31 |
| Существительные и прилагательные | Adjuster Rand:  0.56 0.56 0.57  V measure:  0.01 0.01 0.01  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.33 0.33 0.34 | Adjuster Rand:  0.57 0.57 0.57  V measure:  0.01 0.01 0.02  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.33 0.33 0.33 |
| Существительные | Adjuster Rand:  0.56 0.56 0.57  V measure:  0.01 0.01 0.01  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.33 0.34 0.34 | Adjuster Rand:  0.57 0.57 0.57  V measure:  0.01 0.01 0.01  Mutual info score:  0.01 0.01 0.01  Fowlkes Mallows:  0.33 0.33 0.33 |

Hierarchical algorithm

|  |  |  |
| --- | --- | --- |
| Предобработка\метрика | Евклидова | Косинусная |
| Все части речи | Adjuster Rand:  w:0.57 c:0.54 a:0.40 s:0.34  V measure:  w:0.01 c:0.01 a:0. s:0.  Mutual info score:  w:0.02 c:0.01 a:0. s:0.  Fowlkes Mallows:  w:0.01 c:0.01 a:0. s:0.09 | Adjuster Rand:  c:0.47 a:0.40 s:0.34  V measure:  c:0. a:0. s:0.  Mutual info score:  c:0.01 a:0. s:0.  Fowlkes Mallows:  c:.0.01 a:0. s:0.13 |
| Существительные и прилагательные | Adjuster Rand:  w:0.54 c:0.54 a:0.42 s:0.34  V measure:  w:0.01 c:0. a:0. s:0.  Mutual info score:  w:0.02 c:0. a:0. s:0.  Fowlkes Mallows:  w:0.02 c:0. a:0. s:0.12 | Adjuster Rand:  c:0.49 a:0.43 s:0.34  V measure:  c:0. a:0. s:0.  Mutual info score:  c:0. a:0. s:0.  Fowlkes Mallows:  c:.0.01 a:0.01 s:0.12 |
| Существительные | Adjuster Rand:  w:0.53 c:0.47 a:0.47 s:0.34  V measure:  w:0. c:0. a:0. s:0.  Mutual info score:  w:0.01 c:0. a:0. s:0.  Fowlkes Mallows:  w:0.01 c:0.01 a:0.01 s:0.13 | Adjuster Rand:  c:0.50 a:0.44 s:0.34  V measure:  c:0.01 a:0. s:0.  Mutual info score:  c:0.01 a:0. s:0.  Fowlkes Mallows:  c:.0.01 a:0. s:0.13 |

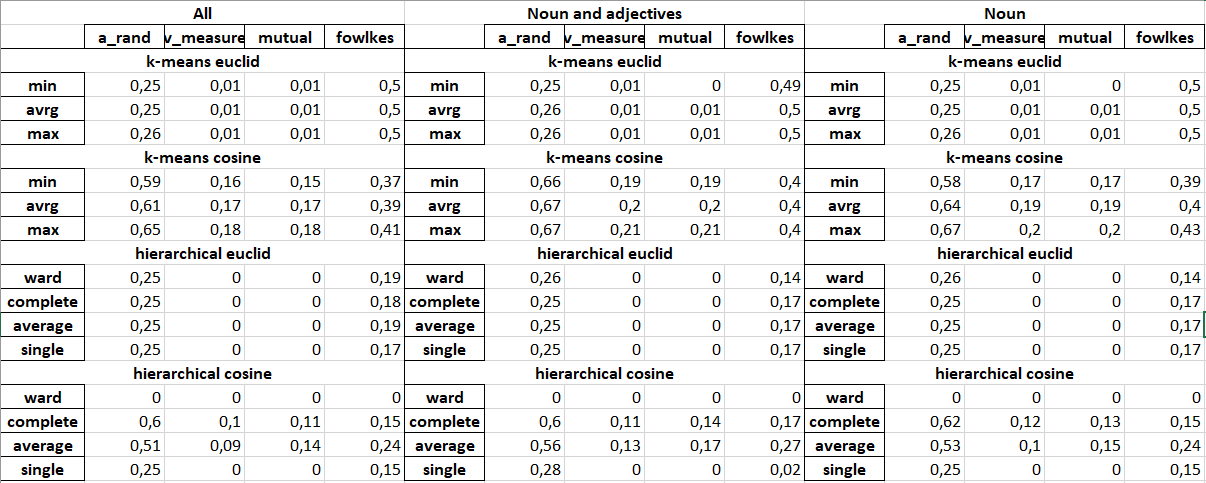
Вывод: на хорошо различимых данных наиболее высокие показатели имеет векторизация с помощью DistilBert и кластеризации k-means на любом наборе частей речи, также близкие результаты демонстрирует DistilBert с иерархической кластеризацией на существительных и прилагательных. На плохо различимых данных высшие показатели получила векторизация BertBaseCased c алгоритмом k-means на существительных и прилагательных; существительных.

1. **20NewsGroups LDA**

****

Вывод: лучшим сочетанием параметров в данном эксперименте стали алгоритм k-means с косинусной метрикой на существительных. Также близкие показатели получил алгоритм иерархической кластеризации на евклидовой и косинусной метрике методами вычисления расстояния average и complete, average соответственно на существительных.

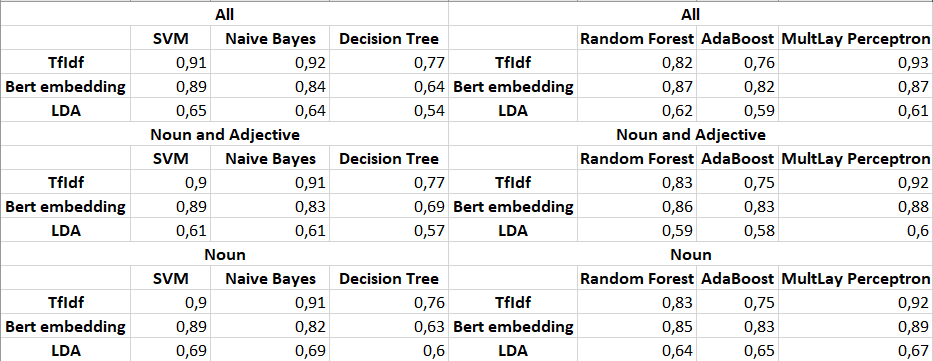
1. **20NewsGroups LSI**

****

Вывод: лучшим сочетанием параметров в данном эксперименте стали алгоритм k-means с косинусной метрикой на существительных и прилагательных; существительных.

Общий вывод по кластеризации: Наиболее высокие показатели по выбранным метрикам в проводимых экспериментах показало сочетание векторизации корпуса с помощью модели LDA и алгоритма k-means с использованием только существительных из исходного корпуса.

**Classification**



Вывод по классификации: в данном эксперименте: наиболее высоким показателем по метрике f1 обладает Multi Layer Perceptron с векторизацией TfIdf; лучшими алгоритмами являются SVM, Naive Bayes и Multi Layer Perceptron; самым лучшей векторизацией является TfIdf.