Walidacja kamienia milowego 1

Zalety:

1. Jasno postawiony, klarowny cel biznesowy.
2. Trafne podejście z usuwaniem duplikatów.
3. Wyrazy takie jak ‘said’, ‘mr’, ‘would’ pojawiają się we wszystkich tematach tekstów. Bardzo dobrze, że potraktowałyście je tak jak stopwordy i wyrzuciłyście.
4. Eleganckie wykresy, dobry pomysł na stworzenie mapy słów.
5. Czytelny plik, każde przekształcenia, operacje są klarownie wytłumaczone. Nie trzeba się niczego domyślać.
6. Plus za wnioski. Widać, że wykonane przekształcenia prowadzą do jakichś konkluzji, nie są to puste komórki kodu, które nie wnoszą nic do postawionego problemu.

Uwagi:

1. Zastanawiamy się, czy w ramach klasteryzacji można w ogóle korzystać z kolumny ‘label’. Naszym zdaniem nie. W końcu jest to uczenie nienadzorowane i formalnie nie powinno korzystać się z etykiet. Podobnie jest z wykresami tworzonymi z podziałem na różne tematy. Rzeczywiście to pozwala lepiej poznać strukturę zbioru danych, jednak w klasteryzacji klastry powinny być raczej wyodrębnione na podstawie zróżnicowania danych (za pomocą jakiegoś modelu), a nie znane już na samym początku.
2. Być może warto usunąć jeszcze jakieś inne, niewiele wnoszące słowa, np. ‘us’.
3. Brak lematyzacji. Usunięto mało istotne słowo ‘said’ dla pewnej grupy tekstów, jednak na wykresie widać, że pozostał rdzeń tego wyrażenia - słowo ‘say’. Lematyzacja rozwiązałaby ten problem.
4. Oprócz analizy występowania pojedynczych słów, warto przeprowadzić podobną analizę dla ich par, tak zwanych ‘bigramów’. Z drugiej strony można to potraktować już jako część Feature Engineeringu.
5. Brak informacji o tym, w jaki sposób następował podział danych na zbiór treningowy, testowy i walidacyjny. Czy był to wybór losowy?

Odpowiedź zespołu walidacji:

1. W notebooku został dodany poniższy komentarz po wyświetleniu zawartości ramki danych: *Jak widać mamy dwie kolumny: Text, która zawiera fragmenty artykułów, oraz Label, która zawiera informacje, do której kategorii został przypisany dany tekst. Kolumna Label oczywiście nie będzie używana w procesie klasyfikacji, ale pozwoli nam na wstępne zapoznanie się z danymi oraz posłuży, aby zweryfikować poprawność klasyfikacji*.
2. Zmodyfikowałyśmy listę angielskich stopwords, dodając wybrane przez nas słowa. Na razie ręcznie, ale pomyślimy nad bardziej automatycznym rozwiązaniem.

3. Lematyzacja przeprowadzona, dane ponownie wyczyszczone.
4. O analizie par i trójek słów myślałyśmy jako część drugiej części projektu, jednak po uwadze od naszego zespołu walidacyjnego dodałyśmy ją już teraz jako wstęp do dalszej części projektu.
5. Podział został wykonany losowo, w oddzielnym pliku. Nie uznałyśmy, że jest to na tyle ważna operacja, żeby znalazła się w pliku budowy. Ale dla naszego zespołu walidacyjnego dodałyśmy na początku komórkę z kodem, który wykonuje podział danych :)

Walidacja kamienia milowego nr 2

Zalety:

1. Dobre podejście z użyciem lematyzacji.
2. Słuszne rozwiązanie bazujące na metodzie tfidf-vectorizer. Rozbicie zdań na pojedyncze słowa może nieść za sobą jakieś ciekawe informacje.
3. Ciekawe podejście z metodą word2vec i GloVe. Wektory słów mogą mieć kluczowe znaczenie w dalszym trenowaniu i uczeniu modelu.
4. Wyniki na zbiorze walidacyjnym nie różnią się szczególnie od wyników uzyskanych przez zespół budowy.

Uwagi:

1. Poniższy fragment kodu nie działa dla danych walidacyjnych:

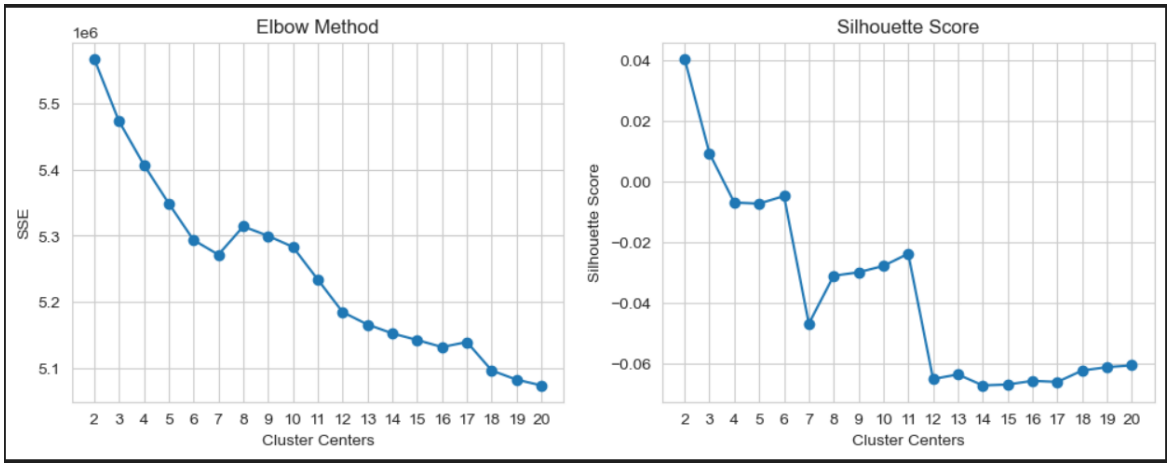
```
# scaler.fit(tfidf_features)
# tfidf_features = scaler.transform(tfidf_features).toarray()

# scaler.fit(sentiment.reshape(-1, 1))
# sentiment = scaler.transform(sentiment.reshape(-1, 1))

# scaler.fit(word2vec_features)
# word2vec_features = scaler.transform(word2vec_features)

# scaler.fit(glove_features)
# glove_features = scaler.transform(glove_features)

# scaler.fit(ngram_matrix)
# ngram_matrix = scaler.transform(ngram_matrix).toarray()

tfidf_features_valid = scaler.transform(tfidf_features_valid)
sentiment_valid = scaler.transform(sentiment_valid)
word2vec_features_valid = scaler.transform(word2vec_features_valid)
glove_features_valid = scaler.transform(glove_features_valid)
ngram_matrix_valid = scaler.transform(ngram_matrix_valid).toarray()
```

⊗ 6.0s

```
features = np.concatenate([tfidf_features, sentiment, word2vec_features, glove_features, ngram_matrix], axis=1)
```

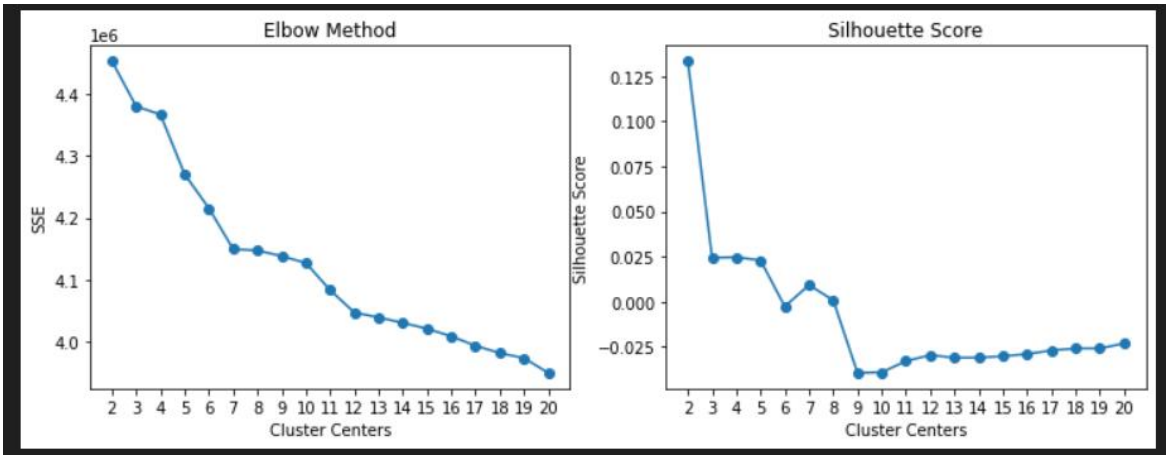
✓ 0.0s

2. Zastanawiamy się, czy łączenie cech utworzonych różnymi metodami w jedną ramkę ma sens. Kolumny uzyskane metodą Tf-Idf niosą ze sobą znacznie mniejszą informację niż kolumny reprezentujące sentyment. Te pierwsze informują wyłącznie o pojedynczych wyrazach, a sentyment niesie informację o całym tekście. Naszym zdaniem warto zbudować oddzielne modele dla każdego z tych podejść i dopiero później porównywać uzyskane na nich wyniki.
3. Bardzo niepokojąco wygląda wykres dotyczący silhouette score. Wartości bliskie 0 sugerują, że coś z danymi jest nie tak.



Wykres dla metody łokcia również w niewielkim stopniu przypomina rzeczywisty,

modelowy łokieć. Przyczyną może być dużo składowych. Po redukcji danych do 400 wymiarów, nadal obserwujemy mało satysfakcjonujące wyniki.



Może warto dokonać kolejnej redukcji danych, a przede wszystkim nie grupować wszystkich metod wykorzystanych w feature engineeringu w jeden zbiór. Może warto też spróbować innych metryk?

4. Standaryzacja zmiennych binarnych może niekoniecznie być dobrym pomysłem. To zawsze i tak będą tylko dwie unikalne wartości, nie ma znaczenia, czy będzie to 0 i 1 czy też 0.2 i 0.3. Tu link do dyskusji na stacku na ten temat: https://stats.stackexchange.com/questions/59392/should-you-ever-standardise-binary-variables

Odpowiedź zespołu walidacji:

Zespół walidacji słusznie zwrócił uwagę, że otrzymane w wyniku etapu 2 wykresy łokcia czy Silhouette odbiegają od modelowych oraz nie przyjmują wartości, które mogłyby sugerować poprawne działanie modelu. Zdecydowałyśmy się więc dodać do testowania algorytmów ramki: word2vec_features, tfidf_features, ngram_features, glove_features - jeszcze niepołączone nowe cechy otrzymane w drugim etapie budowy. Na drugim etapie budowy zdecydowałyśmy, że optymalną liczbą klastrów będzie 5. Dla tej wartości rozpoczniemy klasteryzację. Po uwzględnieniu uwag zespołu walidacyjnego wartość 5 nie będzie już tą dominującą - będziemy sprawdzać wyniki dla różnych wartości.

Walidacja kamienia milowego nr 3

Zalety:

1. Zespół budowy przetestował wiele, bo aż osiem modeli. Każdy z nich został wytrenowany i uruchomiony na sześciu różnych zbiorach danych. Zespół budowy, na skutek naszej sugestii, zrezygnował z łączenia cech uzyskanych z różnych źródeł i wywołał model oddzielnie dla każdego zbioru cech. Dzięki temu wiadomo, który metoda ekstrakcji cech jest najskuteczniejsza w tym zadaniu klasteryzacji.

2. Zespół walidacyjny docenia próby uzyskania więcej niż jednego klastra w metodzie DBSCAN. Jednocześnie zespół budowy podjął słuszną decyzje, że model ten nie działa dla tego zbioru i nie ma sensu dalej go rozważać.
3. Bardzo czytelny kod, dobrze zaagregowane wyniki poszczególnych metryk.

Uwagi:

1. Mimo pokaźnej liczby wykresów nie obserwujemy płynących z nich konstruktywnych wniosków. Przedstawiono wiele wyników poszczególnych metryk, ale zabrakło wyrażonego zdania przez zespół budowy na temat skuteczności i poprawności danej metody klasteryzacji.
2. Wysoki silhouette score dla ramki ngram features oraz metody KMeans, jednak to jedynie efekt wyboru dwóch klastrów. Można byłoby spróbować podejścia, że dla każdej ramki testujemy metryki dla kilku różnych klastrów i następnie wybieramy liczbę klastrów, która ma najwyższy score albo skorzystać z metody łokcia (chyba, że tak zespół budowy zrobił, wtedy jednak warto byłoby to udokumentować w kodzie). Zespołowi walidacji nie podoba się stosowane tu podejście, co do arbitralnego wyboru liczby klastrów dla poszczególnych ramek.
3. Wyniki wszystkich modeli zostały zaprezentowane wyłącznie za pomocą PCA. Być może warto byłoby skorzystać z innych technik wizualizacji, np. t-SNE. Być może wtedy byłoby widać coś więcej.
4. Mimo licznych podsumowań, brakuje informacji o tym, który model został ostatecznie wybrany. Umieszczone zostały wartości poszczególnych metryk, jednak nie wiadomo, który model na ich podstawie jest najlepszy. Nie można patrzeć wyłącznie na wartość jednej z nich. Na przykład Silhouette słabo działa w sytuacji, gdy mamy jeden duży klaster, a pozostałe znacznie mniejsze.
5. Brakuje interpretacji poszczególnych klastrów. Po wybraniu ostatecznego modelu, należy spojrzeć na oryginalną ramkę danych i na podstawie treści artykułu ustalić, co łączy teksty z jednego klastra. Niekoniecznie muszą to być takie tematy, jak na początku.

Porównanie modeli

Zajmiemy się teraz porównaniem wyników poszczególnych modeli na zbiorze treningowym i walidacyjnym.

KMeans

Znacząca różnica wyników metryk dla pierwszych dwóch zbiorów w porównaniu ze zbiorem treningowym i zbiorem walidacyjnym. W przypadku obu tych zbiorów uzyskano znacznie gorsze wyniki każdej metryki na zbiorze walidacyjnym. W kontekście reszty zbiorów różnice są niewielkie. Dla zbioru ngram wyniki Silhouette i Davies-Bouldin są lepsze na danych walidaycjnych.

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.379527	0.812160	3020.862185
1	features_reduced_598	0.355912	0.869804	2673.185093
2	word2vec_features	0.124780	2.328722	99.467361
3	tfidf_features	0.020638	5.373094	5.832340
4	ngram_features	0.594166	0.288293	9.205511
5	glove_features	0.130579	2.288855	91.182583

Zbiór treningowy

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.039851	4.241543	21.054054
1	features_reduced_598	0.036877	4.338805	19.930727
2	word2vec_features	0.125690	2.291017	62.315312
3	tfidf_features	0.018673	5.171165	3.683179
4	ngram_features	0.702342	0.204452	16.352566
5	glove_features	0.133073	2.090382	58.954417

Zbiór walidacyjny

Agglomerative Clustering with Single Linkage

Dla pierwszych dwóch zbiorów ujemny wynik Silhouette na danych treningowych. To znaczy, że obserwacje zostały przydzielone do niewłaściwych klastrów. Wyniki, co ciekawe, generalnie lepsze na zbiorze walidacyjnym.

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	-0.200654	2.164111	0.168211
1	features_reduced_598	-0.202670	2.174128	0.168164
2	word2vec_features	0.156224	0.740267	1.769649
3	tfidf_features	0.005466	0.982213	1.035021
4	ngram_features	0.606327	0.273085	10.536698
5	glove_features	0.366782	0.497036	3.881300

Zbiór treningowy

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.195873	0.688626	2.074324
1	features_reduced_598	0.199948	0.683458	2.106520
2	word2vec_features	0.339978	0.519510	3.590112
3	tfidf_features	0.005560	0.982274	1.034954
4	ngram_features	0.704780	0.167450	46.785047
5	glove_features	0.284295	0.583444	2.807077

Zbiór walidacyjny

Agglomerative clustering with Complete Linkage

Znaczne pogorszenie wszystkich metryk na zbiorze treningowym w przypadku zbiorów features_reduced. Stosunkowo stabilne wyniki dla pozostałych zbiorów. Zbiory skonstruowane metodami word2vec, ngram i glove wychodzą lepiej dla danych walidacyjnych.

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
features_reduced_400		0.368128	0.875062	2716.080319
features_reduced_598		0.344784	0.889184	2585.291610
word2vec_features		0.102221	2.365219	112.019393
ngram_features		0.572442	0.523732	11.913515
glove_features		0.234540	1.978565	24.844636

Zbiór treningowy

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
features_reduced_400		0.031629	3.550358	14.283342
features_reduced_598		0.035815	3.215980	13.937352
word2vec_features		0.137684	1.920808	20.849401
ngram_features		0.631242	0.203146	35.452502
glove_features		0.284653	1.464426	7.326302

Zbiór walidacyjny

Agglomerative clustering with Average Linkage

Podobnie jak w poprzednich metodach obserwujemy znaczący spadek na metryce Silhouette score dla pierwszych dwóch zbiorów danych. Jeszcze większe różnice widzimy dla metryki Calinski-Harabasz score. Jedynie dla ramki danych ngram_features odnotowujemy lepszy wynik na zbiorze walidacyjnym. Jak się jednak później okaże, te obserwacje są mało wartościowe -

rozważamy jedynie dwa klastry, przez co poszczególne metryki mogą prezentować zaburzone wyniki.

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.352609	0.866240	2653.824875
1	features_reduced_598	0.312213	0.857496	2301.509877
2	word2vec_features	0.235252	0.640054	2.367545
3	tfidf_features	0.010802	3.821372	1.221202
4	ngram_features	0.626356	0.263131	11.062535
5	glove_features	0.327626	1.127817	7.784958

Zbiór treningowy

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.122543	0.729476	1.896487
1	features_reduced_598	0.115418	0.737355	1.864201
2	word2vec_features	0.339978	0.519510	3.590112
3	tfidf_features	0.008782	3.505907	1.127287
4	ngram_features	0.850062	0.099698	73.443753
5	glove_features	0.412990	0.449777	4.729709

Zbiór walidacyjny

Agglomerative clustering with Centroid Linkage

Ponownie spadek wartości metryk dla zbiorów features_reduced na danych walidacyjnych. Bardzo stabilne wyniki dla zbioru tfidf, jednak warto podkreślić, że są to dośc słabe wyniki metryk.

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.350696	0.850893	2729.998665
1	features_reduced_598	0.334652	0.907771	2544.133567
2	word2vec_features	0.261657	0.601887	2.677604
3	tfidf_features	0.012843	0.974169	1.051645
4	ngram_features	0.635639	0.255640	11.724677
5	glove_features	0.362951	0.496773	3.885345

Zbiór treningowy

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.131702	0.719265	1.949757
1	features_reduced_598	0.134287	0.725186	1.917155
2	word2vec_features	0.339978	0.519510	3.590112
3	tfidf_features	0.012316	0.974579	1.049545
4	ngram_features	0.850062	0.099698	73.443753
5	glove_features	0.360782	0.477464	4.356562

Zbiór walidacyjny

Agglomerative clustering with Ward’s Linkage

Widzimy jeszcze większy spadek w porównaniu do poprzednich przypadków w metryce Silhouette score dla pierwszych dwóch ramek danych. Gołym okiem widać niestabilność w wynikach. Stabilność jedynie osiągana dla tfidf_features, jednak jest ona obarczona bardzo niskim rezultatem. Ciekawe zachowanie w przypadku ramki ngram_features, jednak liczba klastrów dyskwalifikuje jakiekolwiek konstruktywne wnioski wynikające z rozważanych wyników.

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.356741	0.809601	2784.045186
1	features_reduced_598	0.343767	0.943075	2385.901333
2	word2vec_features	0.114986	2.407567	93.512088
3	tfidf_features	0.010282	7.014632	9.646194
4	ngram_features	0.342491	3.256893	34.147038
5	glove_features	0.127944	2.192827	80.410369

Zbiór treningowy

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.051360	3.786092	23.291391
1	features_reduced_598	0.046079	3.893378	21.257612
2	word2vec_features	0.108158	2.444056	57.871818
3	tfidf_features	0.015279	6.872500	6.181207
4	ngram_features	0.850062	0.099698	73.443753
5	glove_features	0.111911	2.239193	49.978060

Zbiór walidacyjny

Divisive Clustering

Spadek wartości metryk dla zbiorów features_reduced na danych walidacyjnych. Stabilne wyniki w pozostałych przypadkach, jednak jedyny godny zainteresowania wynik Silhouette prezentują dane uzyskane metodą ngram.

	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.248053	1.188782	2350.990014
1	features_reduced_598	0.225958	1.325066	2017.969632
2	word2vec_features	0.114986	2.407567	93.512088
3	tfidf_features	0.010282	7.014632	9.646194
4	ngram_features	0.313987	2.381010	21.533904
5	glove_features	0.117945	2.300591	97.533801

Zbiór treningowy

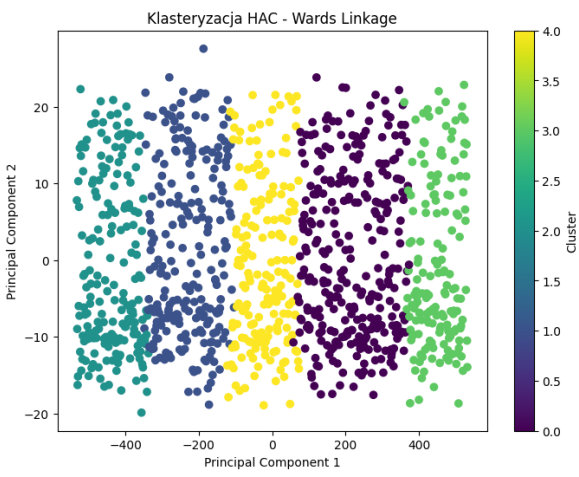
	Data	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	features_reduced_400	0.032380	4.468048	16.284439
1	features_reduced_598	0.028809	4.560082	15.113255
2	word2vec_features	0.108158	2.444056	57.871818
3	tfidf_features	0.015279	6.872500	6.181207
4	ngram_features	0.323550	1.430326	34.665423
5	glove_features	0.105068	2.276591	60.467407

Zbiór walidacyjny

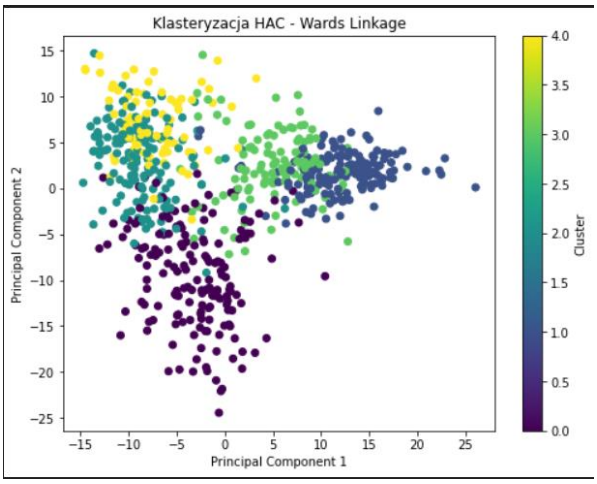
Podsumowując, wyniki uzyskane na danych treningowych i walidacyjnych różnią się znacząco. Jest to dość nietypowe, biorąc pod uwagę, że zbiory zostały podzielone w sposób losowy. Szczególnie niestabilne wydają się modele oparte na zbiorach o nazwie „features_reduced". Naszym zdaniem te zbiory nie powinny być już więcej brane pod uwagę. Dobry wynik Silhouette dla danych uzyskanych metodą ngram wydaje się być związany z niezbalansowanym rozmiarem klastrów tworzonych w tym przypadku. Uzyskujemy bowiem z reguły tylko jeden duży klaster i kilka zawierających pojedyncze obserwacje. Dobry kandydatami na finalne modele są te, które charakteryzują się stabilnością wyników, a jednocześnie zachowują stosunkowo równomierny rozmiar klastrów. Wśród powyższych wyników jest kilka, które osiągają Silhouette na poziomie 0.3 i wyżej, myślę, że warto przyjrzeć się im bliżej (zwłaszcza, jeśli pozostałe warunki też są zachowane). Warto spróbować przyjrzeć się związanym z nimi tekstom i spróbować nadać im jakąś interpretację.

Przyjrzymy się również, czy na zbiorze treningowym i walidacyjnym uzyskujemy podobne wizualizacje. Poniższe wizualizacje dotyczą danych uzyskanych każdą z metod przy użyciu modelu Agglomerative Clustering with Ward Linkage.

Zbiór features_reduced_400

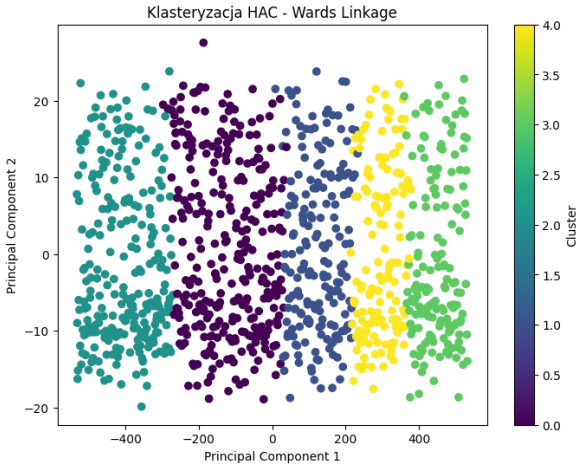


Zbiór treningowy

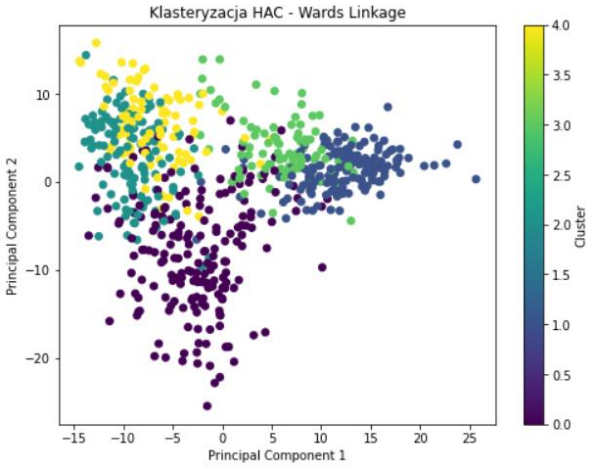


Zbiór walidacyjny

Zbiór features_reduced_598

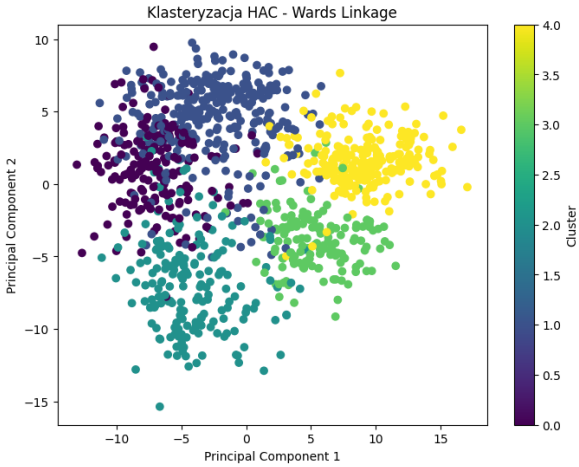


Zbiór treningowy

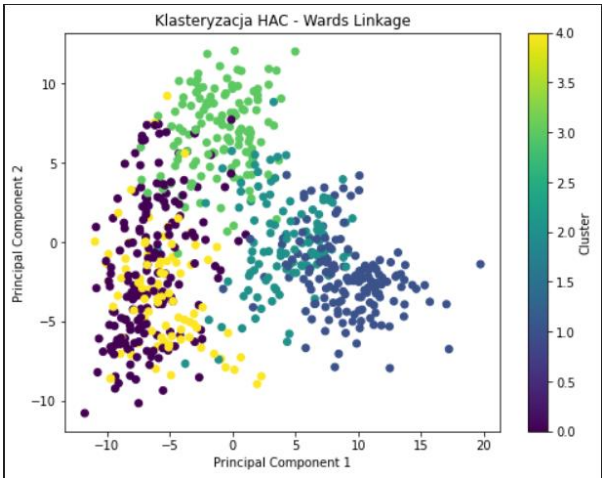


Zbiór walidacyjny

Zbiór word_2vec_features

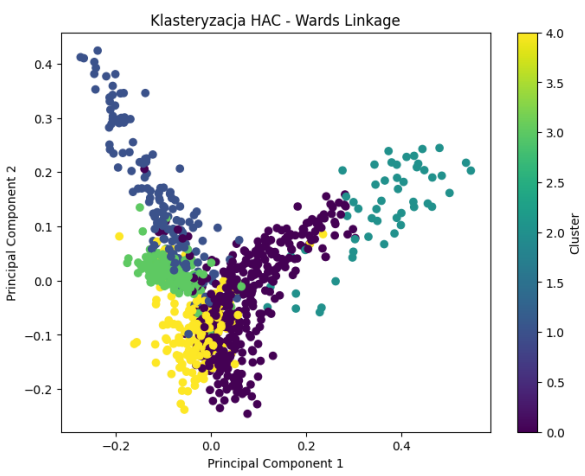


Zbiór treningowy

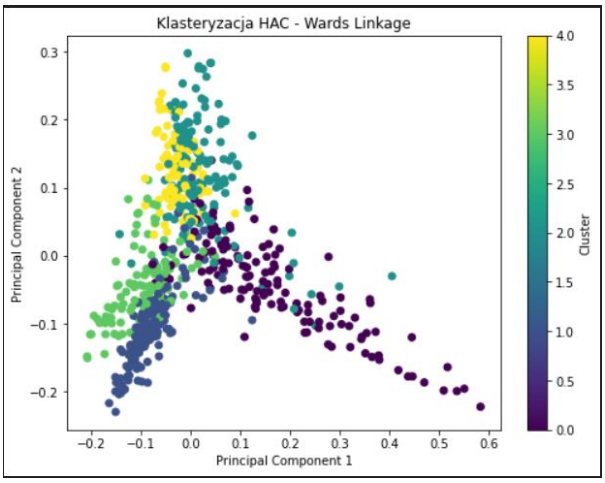


Zbiór walidacyjny

Zbiór tfidf_features

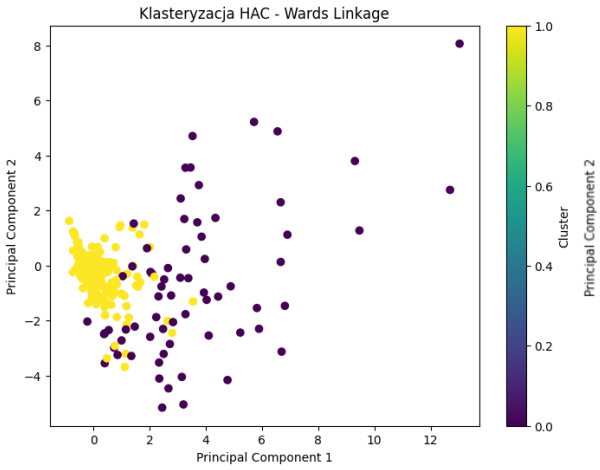


Zbiór treningowy

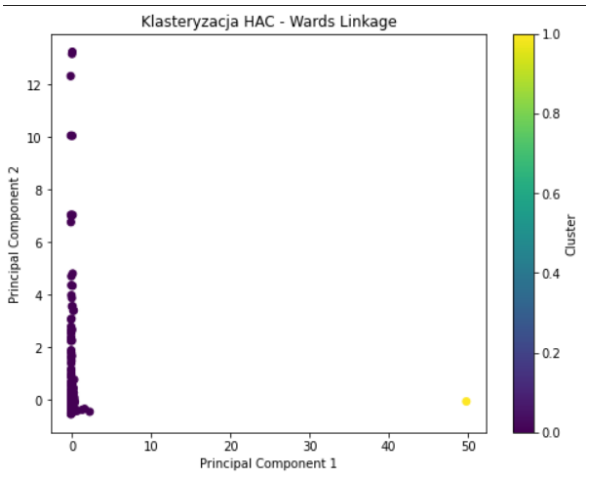


Zbiór walidacyjny

Zbiór ngram_features

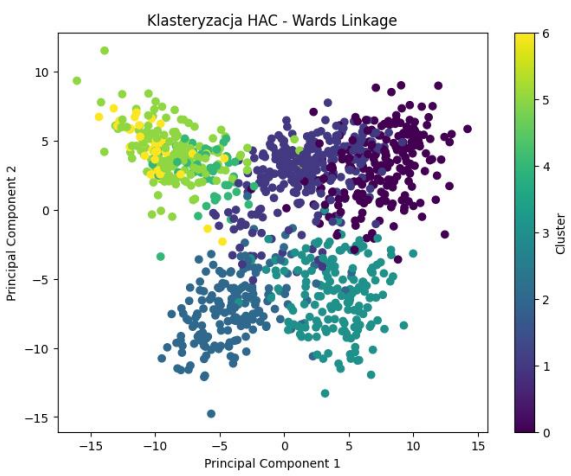


Zbiór treningowy

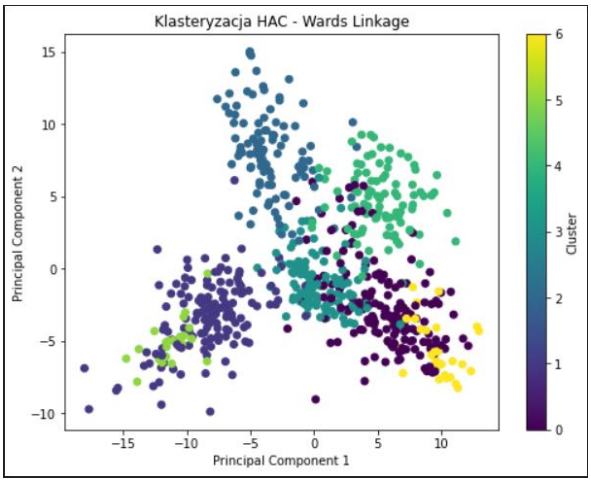


Zbiór walidacyjny

Zbiór glove_features



Zbiór treningowy



Zbiór walidacyjny

Na danych walidacyjnych uzyskaliśmy inne wizualizacje niż na danych treningowych. Mało interesujące wydają się dane uzyskaną metodą ngram, w szczególności dla zbioru walidacyjnego.

Podsumowanie

Zespół walidacyjny nie otrzymał kodu wskazującego jeden z przetestowanych modeli jako najlepszy do realizacji zadania. Nie otrzymano także interpretacji uzyskanych klastrów. Zespołowi budowy przekazane zostały jednak wskazówki odnośnie tego, który model wybrać.