Der Spielraum zwischen „zu wenig“ und „zu viel“

Du, Keli
keli.du@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Ausgangspunkt

Als eine quantitative textanalytische Methode wurde Topic Modeling[1] in den letzten Jahren in Digital Humanities häufig eingesetzt, um zahlreiche unstrukturierte Textdaten zu explorieren. Wenn Topic Modeling verwendet wird, muss man zuerst selbst entscheiden, wie viel Topics trainiert werden sollen. Es ist zwar bekannt, dass die Topic-Anzahl erhebliche Einfluss auf das Topic-Modell hat. Aber es ist nicht so ganz klar, wie groß der Unterschied zwischen den zwei Topic-Modellen ist, wenn man diese zwei Topic-Modelle mit unterschiedlicher Topic-Anzahl auf demselben Korpus trainiert.

In (Wallach et al., 2009) wurde Perplexität als interne Evaluationsmaß des Topic-Modells vorgeschlagen. Ein Topic-Modell wird als ein statistisches Sprachmodell betrachtet. Je niedriger die Perplexitätswerte ist, ist das Modell besser. In (Murphy, 2012, S. 954-955) wurde vorgestellt, dass die Perplexität von LDA-Topic-Modell mit der Erhöhung von Topic-Anzahl reduziert (Abbildung 1). In (Jurafsky & Martin, 2009, S. 43) wurde aber betont, dass die Korrelation zwischen Perplexität und Leistungsfähigkeit des Modells keine Kausalität ist. Deshalb kann eine interne Verbesserung in Perplexität nicht garantieren, dass das Modell bei den externen Aufgaben auf jeden Fall besser funktionieren kann. Eine End-to-End Evaluation (z. B. Dokument-Klassifikation) ist immer notwendig.

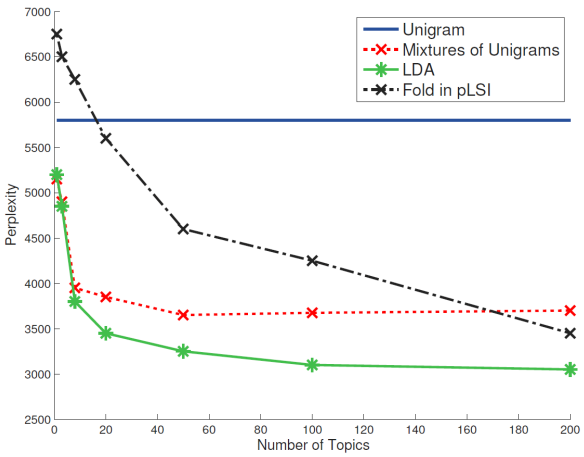

Abbildung 1: Perplexität vs. Topic-Anzahl auf TREC-AP-Korpus[2] (Murphy, 2012, S. 955)

Außerdem, wenn Topic Modeling für Forschung in Digital Humanities eingesetzt wird, interagieren die Benutzer normalerweise direkt mit Topics. Deshalb sind die standardmäßigen internen Evaluationsmethoden[3] für die Evaluation des Topic-Modells nicht ausreichend, weil sie die Qualität bzw. die Interpretierbarkeit der Topics nicht wiederspiegeln können. Über den Einfluss der Topic-Anzahl auf die Interpretierbarkeit der Topics wurde zum Beispiel von Matthew Jockers erklärt, wenn ein Topic-Modell zu viel Topics enthält, könnten die Topics ungenügend semantisch verwandte Wörter enthalten, um sinnvolle und interpretierbare Kontexte/Themen zu bilden. Im Gegensatz dazu, könnte ein Topic-Modell mit zu wenigen Topics dazu führen, dass die Topics zu allgemein sind und sie im ganzen Korpus vorkommen (Jockers, 2013, S. 128).

Aber was heißen eigentlich „zu wenig“ und „zu viel“? Um ein besseres Verständnis von dem Spielraum zwischen „zu wenig“ und „zu viel“ zu bekommen, wurden die vorliegende Untersuchungen durchgeführt. Diese Arbeit konzentriert sich nicht darauf, eine Methode zu finden, die die ideale Topic-Anzahl schätzen kann. Diese Arbeit möchte auch nicht, die Leistungsfähigkeit des Topic-Modells zu evaluieren. Das Ziel der Untersuchung in dieser Arbeit ist den Einfluss der Topic-Anzahl auf das Topic-Modell aus zwei Perspektiven zu verstehen: Topic Modeling basierte Dokument-Klassifikation und Topic-Kohärenz.

Korpus und Tools

Das Korpus der Untersuchung besteht aus 2000 deutschen Zeitungsartikeln zwischen 2001 und 2014[4]. Sie teilen sich in 10 thematische Klassen: „Digital“, „Gesellschaft“, „Karriere“, „Kultur“, „Lebensart“, „Politik“, „Reisen“, „Sport“, „Studium“ und „Wirtschaft“. Jede Klasse enthält 200 Dokumente. Das Korpus enthält insgesamt über 3,4 Millionen Tokens und die durchschnittliche Dokumentlänge ist ca. 1700. Alle Dokumente sind lemmatisiert. Abbildung 2 stellt die Verteilung der Dokumentlänge dar. Die meisten Dokumente enthalten 1400 bis 2000 Lemmata.

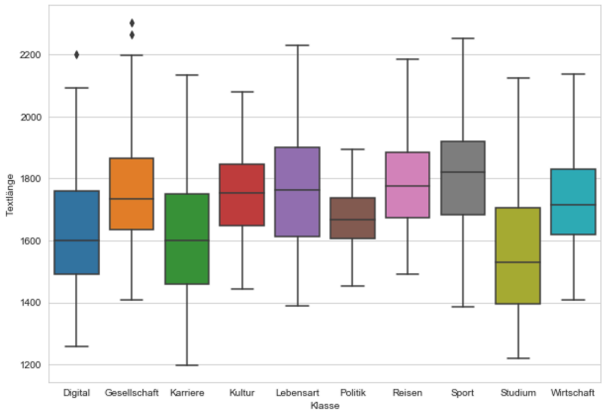

Abbildung 2: Verteilung der Dokumentlänge

Die Topic-Modelle wurden durch MALLET (McCallum, 2002) trainiert. Als Ergebnis bekommt man durch Topic Modeling eine Dokument-Topic-Verteilung und die Topics. In der Dokument-Topic-Verteilung wird jedes Dokument durch einen *n*-dimensionalen Vektor repräsentiert, während *n* die Topic-Anzahl des Topic-Modells ist. Aufgrund der Dokument-Topic-Verteilung wurde die Topic Modeling basierte Dokument-Klassifikation durchgeführt und die Klassifikation erfolgte als 10-fache Kreuzvalidierung mit linearer SVM. Die Topic-Kohärenz wurde durch das Java-Programm Palmetto[5] automatisch berechnet und die erste 10 wichtigste Topic-Wörter wurden für die Berechnung genommen. Das Referenzkorpus für die Berechnung der Topic-Kohärenz ist die lemmatisierte deutschsprachige Wikipedia. In Palmetto wurden mehrere Topic-Kohärenz-Maße implementiert. Für diese Arbeit wurde das Normalised Pointwise Mutual Information (NPMI) basierte Kohärenz-Maß genommen, das in (Aletras & Stevenson, 2013) vorgeschlagen wurde.

Vor der Topic Modeling basierten Dokument-Klassifikation wurde Bag-of-Words (BoW) basierte Klassifikation zuerst durchgeführt, um eine Baseline der Klassifikation zu definieren. Die Tests erfolgten auch als 10-fache Kreuzvalidierung mit linearer SVM[6] , bei welchen der Accurary 0,765 und der F1(Makro)-Wert 0,758 betrug. Eine Baseline des NPMI-Wertes wurde auch definiert. Mit nur einer Iteration wurden zuerst 18 Topic-Modelle auf das Untersuchungskorpus trainiert, die jeweils 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400, 450, 500 Topics enthalten. Dadurch wurden 3150 „Topics ohne Topic Modeling“ erstellt und die NPMI-Werte dieser Topics wurden dann berechnet und der Durchschnittswert ist die NPMI-Baseline: -0,0619. Die Baseline wird durch eine schwarze Linie in den unteren Abbildungen dargestellt.

Die Untersuchungen

Das Ziel der folgenden Untersuchungen ist, den Einfluss der Topic-Anzahl zu überprüfen. Das Setting von Anzahl der Topics war **T** = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400, 450, 500. Für alle anderen Parameter-Einstellungen wurden die vorgegebenen Werte von MALLET genommen. Standard-Stoppwörter wurden vom Korpus entfernt. Aus technischen Gründen, nämlich die zufällige Initialisierung bei der Zuweisung von Topics und Gibbs Sampling, sind zwei Topic-Modelle von einem Korpus nicht völlig identisch, auch wenn das Setting beim Training gleich eingestellt ist. Deshalb können die Ergebnisse der Dokument-Klassifikation und die Topic-Kohärenz von den zwei Modellen unterschiedlich sein. Es wurden deshalb 10 Modelle für jedes Setting trainiert, um den Einfluss von der technischen Seite sichtbar zu machen.

Dokument-Klassifikation: Zuerst wurde der Einfluss von **T** auf die Dokument-Klassifikation mit LDA-Modell untersucht. Abbildung 3 stellt die Verteilungen der Klassifikationsergebnisse dar, die auf 10 Topic-Modelle von jeweiligen Settings basieren. Eine aufsteigende Tendenz ist deutlich erkennbar, wenn **T** von 10 auf 80 erhöht wurde. Eine signifikante weitere Verbesserung der Klassifikation kann in der Abbildung nicht mehr beobachtet werden, wenn **T** von 80 auf 500 erhöht wurde. Die meisten F1-Werte liegen zwischen 0,725 und 0,74. Am besten erzielte die Klassifikation das Accurarcy von 0,759 und den F1-Wert von 0,753. Eine Verbesserung gegenüber der Baseline konnte nicht festgestellt werden. Außerdem ist in der Abbildung zu beobachten, dass es größere Unterschiede unter den 10 Klassifikationsergebnissen gibt, wenn **T** = 10 ist. Diese große Abweichung zeigt, dass die zufällige Initialisierung und Gibbs Sampling eine größere Auswirkung auf das Training des Modells haben, wenn Topic-Modelle mit zu wenig Topics trainiert werden.

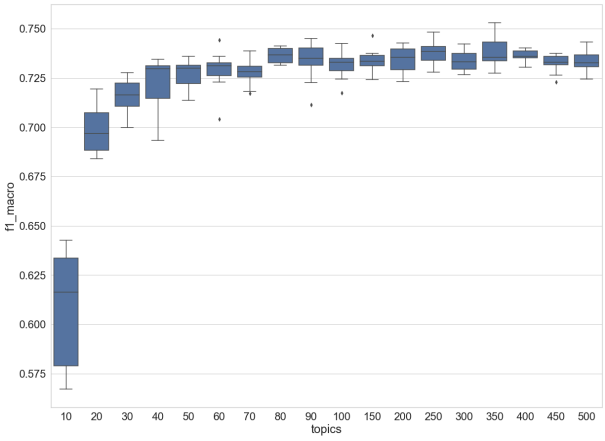

Abbildung 3: F1(Makro)-Werte der Topic Modeling basierten Dokument-Klassifikation im Verhältnis zu Anzahl der Topics

Topic-Kohärenz: Die zweite Untersuchung bezieht sich auf den Einfluss von **T** auf die Kohärenz der Topics. Die Verteilungsdichte der NPMI-Werte wird durch die Violin-Plots in der Abbildung 4 sichtbar dargestellt. Mit der Erhöhung von **T** geht der Median der NPMI-Werte (der weiße Punkt in die Mitte jedes Violin-Plots) unter. Der gesamte Wertbereich der NPMI-Werte ist außerdem breiter geworden, wenn **T** von 10 auf 60 steigt. Der Wertbereich der mittleren 50% der Daten geht mit der Erhöhung von **T** unter und ist hier besonders interessant. Der Bereich verbreitet sich zuerst, wenn **T** von 10 auf 100 steigt. Dann verengt der Bereich sich, wenn **T** von 150 auf 500 steigt.

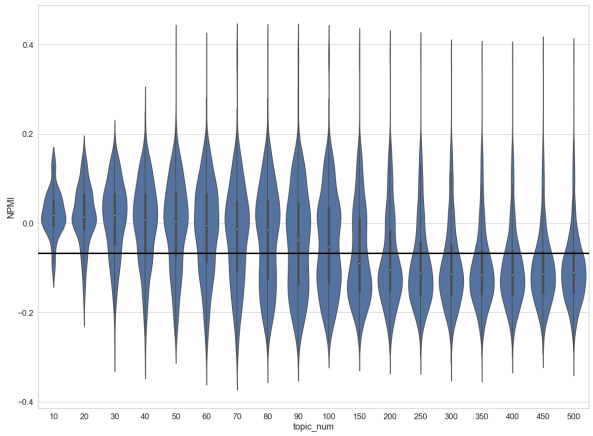

Abbildung 4: NPMI-Werte der Topics im Verhältnis zu Anzahl der Topics

Wenn man die NPMI-Werte der Topics mit der Baseline vergleicht, ist zu sehen, dass man mit der Erhöhung von **T** ständig durch Topic Modeling mehr Topics bekommen kann, deren NPMI-Wert größer als die Baseline ist (Abbildung 5, links). Aber wenn man diese absolute Anzahl normalisiert, also durch die gesamte Anzahl der Topics teilt, ist eine abnehmende Tendenz ganz deutlich erkennbar (Abbildung 5, rechts). Der Anteil von Topics, deren NPMI-Wert größer als die Baseline ist, sinkt von über 90% auf weniger als 30% ab. Das Ergebnis zeigt, dass man mit der Erhöhung von **T** durch Topic Modeling ständig viel mehr nicht kohärente Topics bekommen kann.

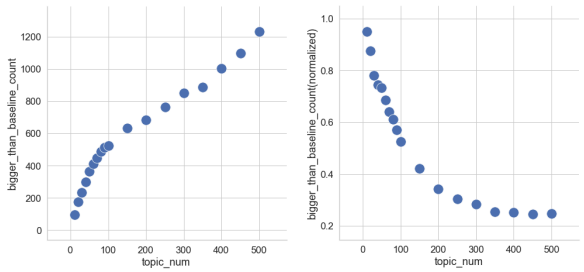

Abbildung 5: Topics, deren NPMI-Wert größer als die Baseline ist (links: absolute Anzahl; rechts: Prozentzahl)

In der Abbildung 4 werden von links nach rechts 10 * **T** , also 100 bis 5000 NPMI-Datenpunkte visualisiert. Um sicherzustellen, dass der Unterschied nicht auf eine ungleiche Anzahl von Datenpunkten zurückzuführen ist, wurde ein weiterer Test gemacht. Es wurden 500 Topic-Modelle à 10 Topics, 50 Topic-Modelle à 100 Topics und 10 Topic-Modelle à 500 Topics trainiert. Danach wurden die NPMI-Werte aller Topics berechnet und visualisiert. In der Abbildung 6 enthalten die drei Violin-Plots jeweils 5000 Datenpunkte. Hier wird eine ähnliche Verteilung wie in der Abbildung 4 beobachtet: Der gesamte Wertbereich verbreitet sich, der Median und der Wertbereich der mittleren 50% der Daten sinken, wenn **T** von 10 über 100 auf 500 erhöht wird.

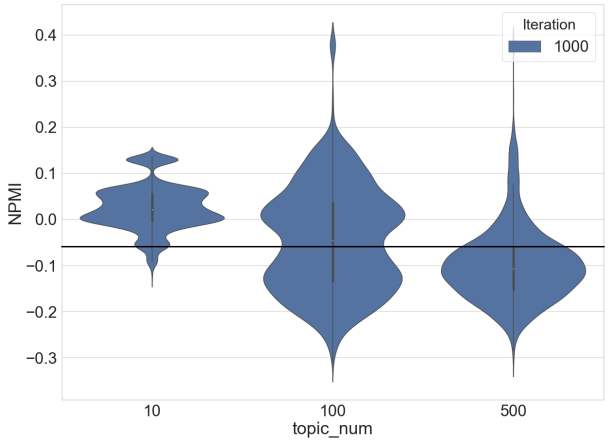

Abbildung 6: NPMI-Wert-Verteilungen von drei Topic-Modelle, die jeweils 5000 Topics enthält

Wenn man die Topics überprüft, sind 10 Topics für 2000 Zeitungsartikel sogar nicht „zu wenig“. In der Tabelle 1 sind die Topics aus einem Topic-Modell mit 10 Topics und sie sind keine allgemeinen Topics, die für Menschen nicht interpretierbar sind. Trotz bestimmter „Geräusche“ (wie z. B. „www“ und „geben“ im Topic 10), können 8 Topics zu den entsprechenden Klassen zugeordnet werden. Auch wenn Topic 6 und 7 zwei Topics sind, in denen vorwiegend Verben gruppiert werden, können sie wegen den Wörtern „kind“, „arbeiten“ und „haus“ mit der Klasse „Lebensart“ oder „Gesellschaft“ verbunden werden.

Nr.	Klasse	NPMI	Top 10 Wörter des Topics
1.	Sport	0,12959	spiel, fußball, spielen, spieler, trainer, tor, wm, fc, minute, verein
2.	Digital	0,06896	internet, facebook, datum, geben, netz, neu, google, online, nutzer, information
3.	Kultur	0.04433	welt, neu, schreiben, film, buch, musik, künstler, spielen, lassen, werk
4.	Wirtschaft	0.03412	euro, prozent, unternehmen, sagen, million, bank, deutsch, firma, neu, geben
5.	Politik	0.01903	sagen, deutsch, geben, krieg, regierung, russisch, neu, präsident, polizei, staat
6.	???	0.01083	sagen, geben, kind, online, wissen, mal, sehen, finden, einfach, arbeiten
7.	???	0.00799	sagen, stehen, haus, sehen, geben, mal, fahren, paar, sitzen, liegen
8.	Politik	0.00109	politisch, neu, politik, lassen, geben, kirche, stehen, politiker, spd, öffentlich
9.	Studium	-0.00441	universität, sagen, hochschule, uni, studium, schule, deutsch, geben, prozent, studieren
10.	Reisen	-0.06509	essen, lassen, restaurant, geben, www, hotel, tier, küche, kochen, liegen

Tabelle 1. 10 Beispieltopics aus einem 10-Topic Topic-Modell

In der Tabelle 2 sind drei Gruppen von Beispieltopics aus drei Topic-Modellen, die jeweils 10, 100, 500 Topics enthalten. Da die meisten Wörter in Topic 1a, 2a und 3a (auch 1b, 2b und 3b) gleich sind, kann festgestellt werden, dass die sinnvollen Topics mit der Erhöhung von **T** , statt sich in mehreren nicht kohärente Topics aufzulösen, kohärent bleiben können. Durch die Erhöhung von **T** werden eher weitere spezifische Topics produziert, wie z.B. „fußball, verein, fc, fan, stadion, bayern, bundesliga, spieler, trainer, liga“ oder „spielerin, birgit, frauenfußball, neid, tor, länderspiel, trabant, wm, sobiech, dfb“.[7]

Nr.	Topics-Anzahl (T)	NPMI	Top 10 Wörter des Topics
1a.	10	0,12959	spiel, fußball, spielen, spieler, trainer, tor, wm, fc, minute, verein
1b.	10	0,06896	internet, facebook, datum, geben, netz, neu, google, online, nutzer, information
2a.	100	0,10963	spieler, spielen, spiel, fußball, trainer, wm, ball, team, deutsch, tor
2b.	100	0,11812	facebook, internet, netz, nutzer, netzwerk, google, twitter, sozial, information, datum
3a.	500	0,10963	spielen, spieler, spiel, fußball, trainer, wm, ball, tor, deutsch, team
3b.	500	0.09401	facebook, netzwerk, internet, netz, nutzer, sozial, twitter, google, freund, information

Tabelle 2: Drei Gruppen von Beispieltopics aus drei Topic-Modellen

Fazit

Die vorliegenden Untersuchungen haben den Spielraum im Sinne von Topic-Anzahl bei Topic Modeling aus zwei Perspektiven eingegrenzt, nämlich Dokument-Klassifikation und Topic-Kohärenz. Angesichts der Untersuchung ist es festzustellen, dass man vermeiden sollte, ein Topic-Modell mit zu wenig Topics zu trainieren, wenn man eine bessere Topic Modeling basierte Dokument-Klassifikation sichern möchte. Ein Topic-Modell mit hoher Topic-Anzahl zu trainieren kann auch mehr kohärente Topics erzielen. Aber gleichzeitig muss man mit noch mehr nicht kohärenten Topics kämpfen. Deshalb ist es notwendig, die Topic-Kohärenz nach Topic Modeling zu berechnen, um die kohärenten und die nicht kohärenten Topics zu unterscheiden. Am Ende muss noch betont werden, dass das Ergebnis abhängig von Untersuchungskorpus sein könnte. Deshalb ist es geplant, die gleiche Untersuchung auf anderen Korpora (z. B. statt Zeitungsartikeln eine Sammlung von literarischen Texten für die Untersuchung zu nehmen) in der Zukunft durchzuführen.

Fußnoten

1. Topic Modeling ist eine Reihe von Algorithmen. Da das Latent Dirichlet allocation (LDA)-Modell am meisten verbreitetes Topic-Modell ist, bezieht „Topic Modeling“ und „Topic-Modell“ in dieser Arbeit sich nur auf LDA.
2. http://www.daviddlewis.com/resources/testcollections/trecap/
3. Hier beziehen die internen Evaluationsmethoden sich nicht nur auf die Perplexität, sondern auch auf die Methoden, die in (Deveaud, SanJuan, & Bellot, 2014); (Arun, Suresh, Veni Madhavan, & Narasimha Murthy, 2010); (Cao, Xia, Li, Zhang, & Tang, 2009) und (Griffiths & Steyvers, 2004) vorgeschlagen wurden
4. Das Korpus ist eine private Textsammlung, die leider nicht veröffentlicht werden kann. Mit MALLET wurde das Korpus importiert und in eine MALLET-Datei umwandelt, die hier verfügbar ist: https://www.dropbox.com/s/jpfhmtneu8q352z/Zeit_10_Klasse_lemma.mallet?dl=0
5. http://aksw.org/Projects/Palmetto.html
6. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
7. „birgit“, „trabant“ und „sobiech“ sind drei Namen, die mit dem Thema Frauenfußball verbunden sind. Birgit Prinz und Anne Trabant sind zwei ehemalige deutsche Fußballspielerinnen. Gabriele Sobiech ist die Autorin vom Buch „ *Spielen Frauen ein anderes Spiel?:- Geschichte, Organisation, Repräsentationen und kulturelle Praxen im Frauenfußball*“

Bibliographie

Aletras, N. / Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* (pp. 13-22).

Arun, R. / Suresh, V., / Veni Madhavan, C. E. / Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Hrsg.), Advances in Knowledge Discovery and Data Mining (Bd. 6118, S. 391–402). https://doi.org/10.1007/978-3-642-13657-3_43

Cao, J. / Xia, T. / Li, J. / Zhang, Y. / Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Deveaud, R. / SanJuan, E. / Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document Numérique, 17(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Griffiths, T. L. / Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.

Jurafsky, D. / Martin, J. H. (2009): *Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed* . Upper Saddle River, N.J., London: Pearson Prentice Hall (Prentice Hall series in artificial intelligence).

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (13.12.2019).

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective* . MIT press.

Wallach, H. M. / Murray, I. / Salakhutdinov, R. / Mimno, D. (2009, Juni): Evaluation methods for topic models. In: *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1105-1112), ACM.

TREC-AP-Korpus: http://www.daviddlewis.com/resources/testcollections/trecap/ (14.12.2019)