Text- und Webmining
GrapplingInsider

Heiko Raible (769082), Leopold Groznova (762049)

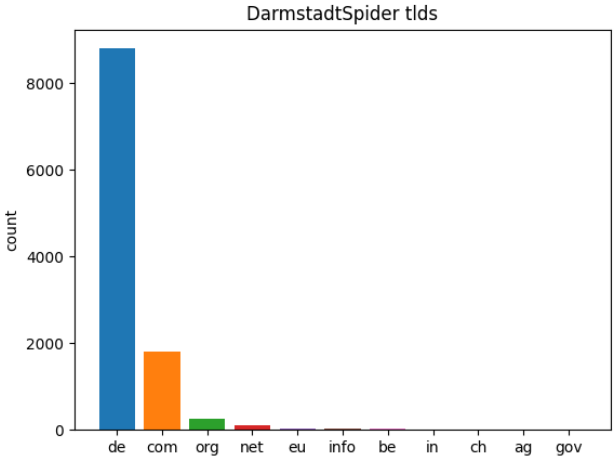
15. Januar 2023

Inhaltsverzeichnis

1 Praktikum 1: Crawling, Linkextraktion und Content-Extraktion

1.1 DarmstadtSpider Analyse

Top-Level Domains



Die mit Abstand häufigste Top-Level Domain (TLD) der verlinkten Webseiten ist 'de', was bei einer deutschen Webseite zu erwarten war. Am Zweithäufigsten mit etwa einem viertel dessen Häufigkeit steht 'com', was ebenfalls nicht verwunderlich ist, da diese TLD international, insbesondere in der westlichen Welt, am meisten vertreten ist.

Links

Folgende ausgehende URLs, welche von mehreren Webseiten-Bereichen verlinkt wurden, sind uns besonders aufgefallen:

aka55plus.de
darmstadt.de/leben-in-darmstadt/soziales-und-gesellschaft/seniorinnen-und-senioren/
darmstadt.de/leben-in-darmstadt/bildung/aus-und-weiterbildung/

⇒ Akademie 55plus Darmstadt e.V. (Aka 55plus) ist ein Verein für Bildung für Menschen ab dem 55. Lebensjahr. Somit passen die Bereiche `soziales-und-gesellschaft/seniorinnen-und-senioren` und `bildung/aus-und-weiterbildung` sehr gut.

cbf-da.de
darmstadt.de/leben-in-darmstadt/soziales-und-gesellschaft/menschen-mit-behinderung/
darmstadt.de/leben-in-darmstadt/mobilitaet-und-verkehr/barrierefrei/

⇒ Club Behinderter und ihrer Freunde in Darmstadt und Umgebung e.V. (CBF) setzt sich für Menschen mit Behinderung ein. Auch hier passen die Bereiche `soziales-und-gesellschaft/menschen-mit-behinderung` und `mobilitaet-und-verkehr/barrierefrei` sehr gut.

gesetze-im-internet.de
darmstadt.de/leben-in-darmstadt/soziales-und-gesellschaft/frauen/aufgaben-und-ziele/
darmstadt.de/leben-in-darmstadt/umwelt/laerm/fluglaerm-in-darmstadt/laermschutzbereich

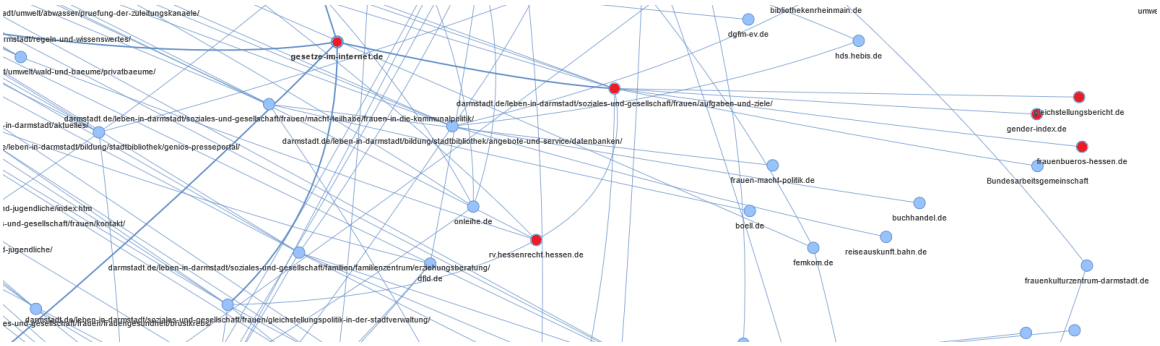
darmstadt.de/leben-in-darmstadt/soziales-und-gesellschaft/frauen/gleichstellungspolitik
darmstadt.de/leben-in-darmstadt/soziales-und-gesellschaft/frauen/gewaltschutz/
darmstadt.de/leben-in-darmstadt/mobilitaet-und-verkehr/radfahren-in-darmstadt/regeln-u

⇒ All diese Bereiche, egal ob es um Frauenrechte, Fluglärm oder Radfahren geht, verweisen auf rechtliche Informationen auf der Webseite gesetze-im-internet.de des Bundesministeriums der Justiz und des Bundesamtes für Justiz.

Weitere Analyse

Als weitere Analyse entschieden wir uns für eine Graphrepräsentation mittels `networkx` und `pyvis`.

Bei genauerer Betrachtung zeigte uns diese Darstellung die eben gelisteten Informationen, konnten dann aber auch weiter verfolgt werden, um zum Beispiel von `gesetze-im-internet.de` über `darmstadt.de/leben-in-darmstadt/soziales-und-gesellschaft/frauen/aufgaben-und-ziele` auf weitere Rechts-bezogene Seiten wie `rv.hessenrecht.hessen.de` oder Frauen-bezogene Seiten wie `gleichstellungsbericht.de`, `gender-index.de` oder `frauenbueros-hessen.de` zu gelangen.



1.2 GrapplingInsiderSpider

Wir starteten auf der Seite `https://grapplinginsider.com/news`, die eine Auflistung der neusten Artikel beinhaltet und bewegten uns von dort in jeden einzelnen Artikel, sowie über den Next Button auf die nächste Seite der neusten Artikel, bis alle Artikel der Webseite gecrawled waren. Je Artikel crawlten wir URL, Kategorien, Titel, Text und Links, letztere als Ziel-URL und zugehörigem Text.

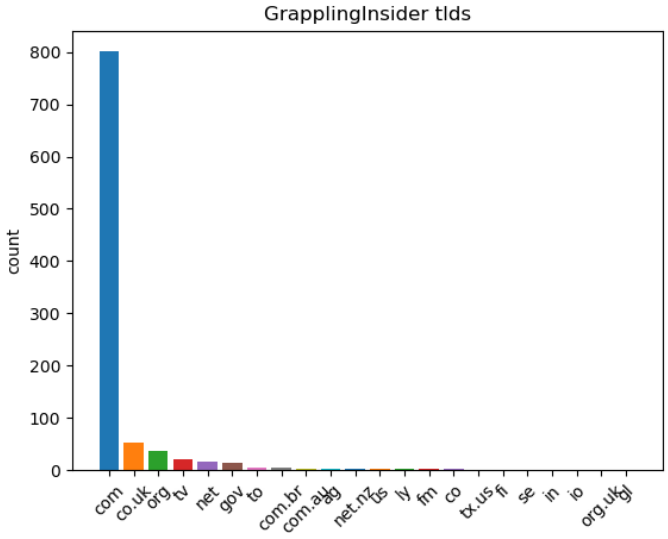
Siehe `GrapplingInsiderSpider.py`.

1.3 XPath und Tupel

```
categories_xpath = response.xpath("//a[@rel='category tag']")
title_xpath = response.xpath("//header/h1[@itemprop='headline']")
text_xpath = response.xpath("//div[@itemprop='articleBody']")
```

1.4 GrapplingInsiderSpider Analyse

Top-Level Domains



Die mit Abstand häufigste Top-Level Domain (TLD) der verlinkten Webseiten ist 'com', was bei einer amerikanischen Webseite zu erwarten war. Selten kamen weiterhin 'co.uk' als weitere englische TLD und die üblichen 'org' für Organisationen, 'tv' welche häufig für Streaming Webseiten genutzt wird, 'net' als häufige generische TLD und 'gov' für Regierungen vor. Die Verteilung ist noch stärker auf die häufigste TLD konzentriert als in den Darmstadt Daten, da die international häufige 'com' TLD hier direkt mit an Platz eins ist, statt separat neben 'de'.

Links

Folgende ausgehenden URLs, welche von mehreren Seiten verlinkt wurden, sind uns besonders aufgefallen:

bbc.co.uk
grapplinginsider.com/uaejjf-postpones-all-bjj-events-over-coronavirus-fears/
grapplinginsider.com/coronavirus-outbreak-cancels-adcc-mongolia/
grapplinginsider.com/british-judo-black-belt-found-guilty-of-murder/
grapplinginsider.com/demi-lovato-just-got-a-stripe-on-her-blue-belt-and-its-awesome/

foxbusiness.com
grapplinginsider.com/former-nypd-detective-pushes-for-brazilian-jiu-jitsu-training-for
grapplinginsider.com/love-him-or-hate-him-you-really-should-not-be-excited-to-see-pres

nytimes.com
grapplinginsider.com/the-new-york-times-investigates-fight-sports-sexual-assault-alleg
grapplinginsider.com/watch-backyard-fight-club-streetbeefs-holds-its-first-grappling-m
grapplinginsider.com/khabib-to-mcgregor-you-are-a-rapist/

usatoday.com
grapplinginsider.com/love-him-or-hate-him-you-really-should-not-be-excited-to-see-pres
grapplinginsider.com/ufc-targets-askren-maia-matchup/

winknews.com
grapplinginsider.com/the-new-york-times-investigates-fight-sports-sexual-assault-alleg
grapplinginsider.com/roberto-cyborg-abreu-announces-new-guidelines-to-address-miscondu
grapplinginsider.com/roberto-cyborg-abreu-and-vagner-rocha-issue-statements-regarding-

⇒ Diese Artikel behandeln News die es in die Mainstream Nachrichten geschafft haben.

healthline.com
grapplinginsider.com/rib-injuries-in-bjj-causes-diagnosis-and-prevention/
grapplinginsider.com/high-rollerz-bjj-and-the-cannabis-community/

ncbi.nlm.nih.gov
grapplinginsider.com/best-supplements-to-help-you-sleep-after-bjj/
grapplinginsider.com/focus-on-sleep-not-supplements/
grapplinginsider.com/bjj-low-testosterone-test-lets-get-checked/

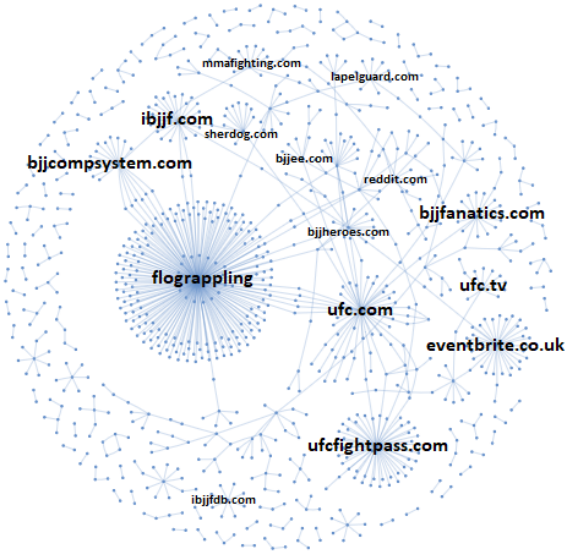
⇒ Diese Artikel behandeln Gesundheit und Medizin.

mginaction.com
grapplinginsider.com/marcelo-garcia-set-to-release-butterfly-guard-instructional/
grapplinginsider.com/5-reasons-why-marcelo-garcia-is-the-greatest-of-all-time/

⇒ Diese Artikel behandeln einen bekannten Wettkämpfer und verweisen auf dessen Webseite.

Weitere Analyse

Auch hier entschieden wir uns wieder für eine Graphrepräsentation mittels `networkx` und `pyvis`.



Hier sieht man klare Cluster von Artikeln um Webseiten großer, bekannter Organisationen. Die größten sind die Webseiten die Events streamen (`flograppling`, `ufc`), die großen Veranstalter und Veranstaltungsaggregatoren (`ufc`, `ibjjf`, `eventbrite`), `bjjfanatics` welche Lehrvideos verkauft, Foren (`sherdog`, `reddit`) und weitere News Webseiten (`bjjheroes`, `bjjee`, `mmafighting`).

2 Praktikum 2 - Teil I: Informationsextraktion und Textzerlegung

2.1 HANA VM

```
connection = dbapi.connect('192.168.56.102', 39041, 'SYSTEM', 'Password1')
```

2.2 Tabellen

```
CREATE TABLE GRAPPLING_INSIDER_CONTENT(
    url VARCHAR(300) PRIMARY KEY,
    title VARCHAR(300),
    text NCLOB MEMORY THRESHOLD 1000
)
CREATE TABLE GRAPPLING_INSIDER_CATEGORIES(
    url VARCHAR(300),
    category VARCHAR(30)
)
CREATE TABLE GRAPPLING_INSIDER_EXTERNAL_LINKS(
    url VARCHAR(300),
    external_link_url NCLOB MEMORY THRESHOLD 1000,
    external_link_text VARCHAR(700)
)
```

2.3 Füllen der DB

```
INSERT INTO GRAPPLING_INSIDER_CONTENT (url, title, text) VALUES (?,?,?)
INSERT INTO GRAPPLING_INSIDER_CATEGORIES (url, category) VALUES (?,?)
INSERT INTO GRAPPLING_INSIDER_EXTERNAL_LINKS (url, link_url, link_text) VALUES (?,?,?)
```

2.4 Prüfen

Alle Daten wie erwartet in der Datenbank.

2.5 Text-Index

```
CREATE FULLTEXT INDEX "GRAPPLING_INSIDER_INDEX"
ON "SYSTEM"."GRAPPLING_INSIDER_CONTENT" ("TEXT")
CONFIGURATION 'LINGANALYSIS_FULL'
ASYNC LANGUAGE DETECTION ('en')
TEXT ANALYSIS ON
```

2.6 Text-Index prüfen

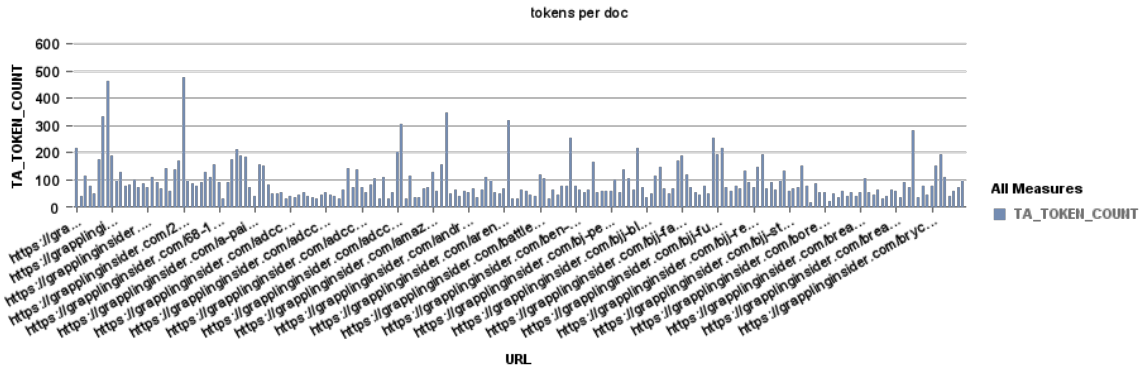
Es standen drei Indexarten zur Auswahl:

- LINGANALYSIS_BASIC: Tokenisierung mit unter anderem Erkennung der Sprache, Normalisierung (bei uns nur forced lower-case), Paragraphen- und Satzindex, sowie Offset im Dokument.
- LINGANALYSIS_STEMS: Hier werden noch die Lemmatisierungen der Tokens hinzugefügt, was bei uns nur mangelhaft funktionierte. Im Satz den wir uns angeschaut haben war nur das Entfernen von `s` am Wortende bei Mehrzahl (zB `friends` zu `friend`) aufzufinden, alles andere hat gefehlt (zB `confined` zu `confine`).
- LINGANALYSIS_FULL: Zuvor wurde der Token Typ nur zwischen `punctuation` und allen anderen unterschieden. Bei dieser Einstellung kommt die restliche PoS Analyse dazu. Dies funktionierte größtenteils gut.

Die Lemmatisierung (TA_STEM) war in unserem Fall nicht nutzbar und wir wissen nicht woran es lag, bzw. ob wir darauf hätten Einfluss nehmen können. Die PoS Analyse (TA_TYPE) war zufriedenstellend und könnte nur noch einmal mit Kontext wichtigen Bedeutungen korrigiert werden, wie man bei späteren Analysen sieht. Bei unseren Daten wären das zum Beispiel Namen wie `Jiu Jitsu` (Sport) oder dem `Gi` (Kleidung) als `proper name` zu definieren, statt wie hier häufig als Adjektive.

2 Praktikum 2 - Teil II: Reporting auf zerlegten Texten

2.1 Worthäufigkeiten (Nomen) pro Dokument

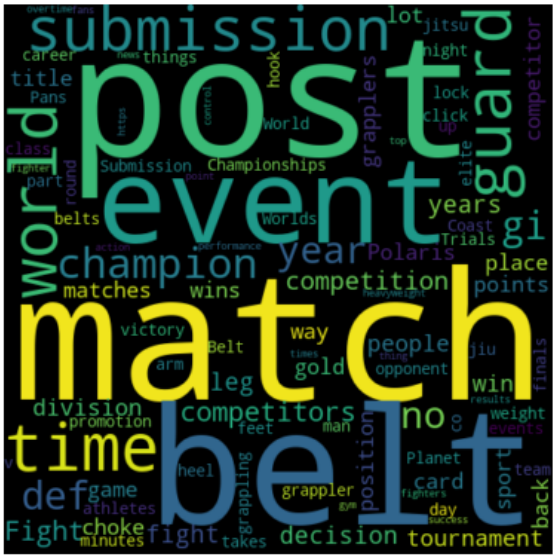


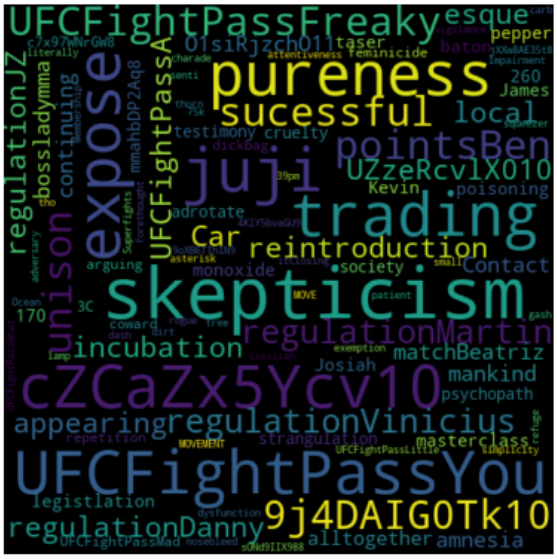
2.2 Lexikon

```
lexicon length: 26936 tokens
filtered lexicon length: 25525 tokens
average document length: 525 tokens
average sentence length: 23 tokens
```

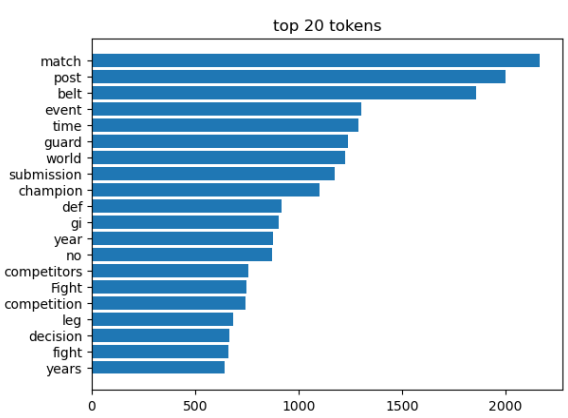
2.3 Verteilung der Worthäufigkeiten

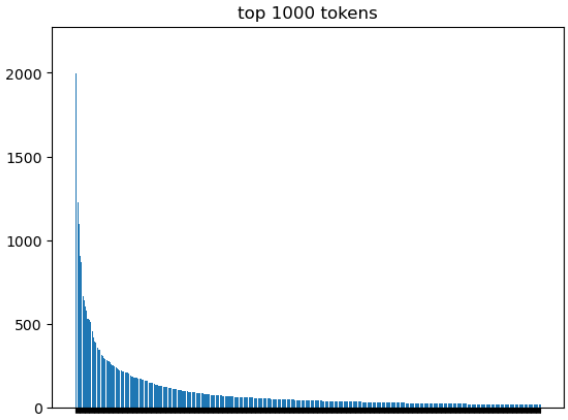
Auf Nomen beschränkt:





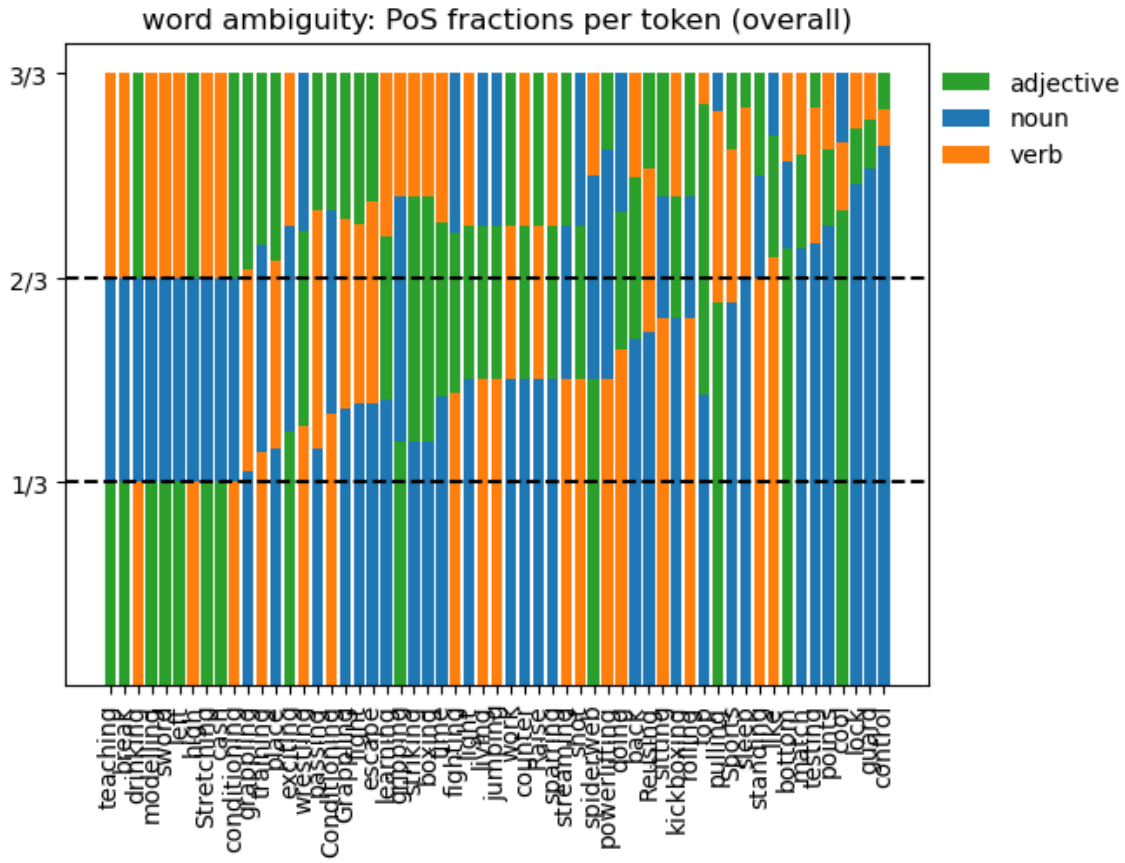
Die top 100 Tokens der WordCloud sehen sehr wie erwartet aus, ohne Überraschungen. Die bottom 100 Tokens sind großteils Unsinn und besteht aus Worten die nur ein mal vorkommen. Es gibt deutlich mehr als 100 Tokens die nur ein mal vorkommen, somit ist die Auswahl auch sehr willkürlich





Die top 20 tokens sind auch nicht überraschend. Bei einer Darstellung der top 1000 tokens erkennt man, dass das Zipfsche Gesetz grob gelten könnte.

2.4 Mehrdeutigkeit



In diesem Plot sind die mehrdeutigsten Worte dargestellt. `teaching` wird beispielsweise genau gleich oft als Verb, Adjektiv und Nomen gebraucht. `control` hingegen wird über $\frac{5}{6}$ mal als Nomen gebraucht und nur etwa jeweils $\frac{1}{12}$ mal als Adjektiv oder Verb.

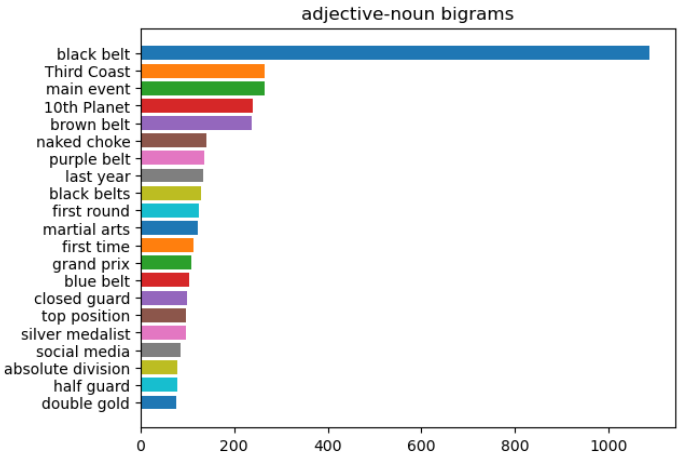
2.5 Weitere Statistiken und Visualisierungen

```
url count: 1375
categories: ['ADCC News', 'Academies', 'BJJ Culture', 'BJJ History',
        'BJJ Injury', 'BJJ News', 'COVID', 'Celebrity', 'Conditioning',
        'Endurance', 'Featured', 'Fitness', 'Health', 'IBJJF News',
        'Interview', 'Judo News', 'MMA News', 'Media', 'Opinion',
        'Preview Events', 'Review Events', 'Reviews', 'Technique',
        'UK BJJ News', 'Uncategorized', 'Video', 'Wellbeing',
        'White belt', "Women's BJJ News", 'Wrestling News']
categories count: 30
```

Als weitere Statistiken haben wir uns die Anzahl an Artikeln und die Kategorien samt Anzahl derer ausgeben lassen.

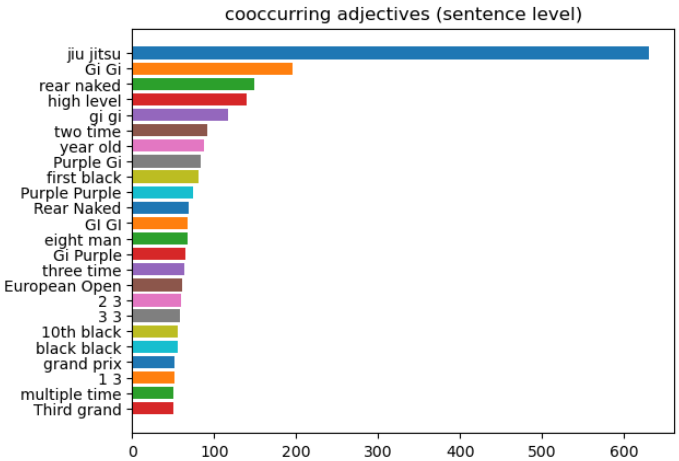
3 Praktikum 3 - Teil I: Fortgeschrittenes Reporting und Dokumentähnlichkeit auf zerlegten Texten (SQL)

3.1 Bigramme



Dass der Schwarzgurt alle anderen Adjektiv-Nomen Bigramme so stark abhängt war etwas überraschend. Auch andere Gurte sind vertreten (`black/brown/purple/blue`), interessanterweise in Reihenfolge vom höchstgradierten bis niedrigstgradierten, mit Ausnahme des weißen Gurtes, welcher der Anfänger Gürtel ist und in absoluter Anzahl die meisten Teilnehmer ausmacht. Das Interesse scheint demnach stark auf den Wettkämpfen und sozialem Drama der `black belt` Klasse zu liegen. Die Adjektive der `main event, first round, first time` waren zu erwarten. Das Erste oder der Hauptteil sind von besonderem Interesse. `10th Planet` ist eine große Organisation im No-Gi Submission Grappling, worauf die Webseite einen Fokus legt gegenüber Gi Grappling. `naked choke, closed guard, top position, half guard` sind zentrale Positionen und Techniken aus dem BJJ.

3.2 Co-occurences



`jiu jitsu, Gi Gi, gi gi, Purple Gi, Purple Purple, GI GI, 2 3, 3 3, black black, 1 3` ist unsinniger Output, der aus einer falschen TA_TYPE bestimmung stammt.

`rear naked, Rear Naked` ist wieder Teil der Technik "Rear Naked Choke", wobei die adjektiv Kombination `rear naked` alleinstehend im Grappling Kontext lustige Bilder hervorruft. `high level, two time, year old, first black, three time, European Open, grand prix, multiple time, third grand` sind alle sehr passend zu einem Wettkampf Kontext. Die Seite scheint wirklich besonders darüber zu berichten.

3.3 tf-idf

```
word: box, url: https://grapplinginsider.com/the-bjj-box-why-you-should-buy-one/
tf-df: 48.17, tf: 29, idf: 1.66

word: rib, url: https://grapplinginsider.com/rib-injuries-in-bjj-causes-diagnosis-and-
tf-df: 47.46, tf: 30, idf: 1.58

word: pounds, url: https://grapplinginsider.com/nogi-worlds-competitor-list-and-previe
tf-df: 41.02, tf: 30, idf: 1.37
```

Das sind alles Worte die in den Dokumenten sehr häufig vorkommen (tf), aber im Corpus nahezu garnicht (idf). Die Dokumente handeln sehr zentral von diesen Worten. Bei letzterem nur indirekt, aber `pound` wird in dieser 'competitor list' häufig gelistet. Dass aufgrund des *ntn* Schemas nach SMART Notation keine Normalisierung stattfand war nicht von besonderer Bedeutung, da die Dokumente nicht viel länger sind als andere.

3.4 similarity

```
def get_similar(self, target_url, n=10, mode="cos"):  # modes = ["scalar", "cos"]
    command = """
    SELECT URL, TA_TOKEN
    FROM "$TA_GRAPPLING_INSIDER_INDEX"
    WHERE TA_TYPE = 'noun'
    ORDER BY URL ASC
    """
    self.cursor.execute(command)
    # create data
    vecs_dict = {}
    for res in self.cursor:
        url = res[0]
        noun = res[1]
        if url not in vecs_dict:
            vecs_dict[url] = {}
        if noun in vecs_dict[url]:
            vecs_dict[url][noun] += 1
        else:
            vecs_dict[url][noun] = 1
    data = pd.DataFrame(vecs_dict)
```

```
    data = data.fillna(0)
    data = data.astype(int)
    data = data.transpose()
    # get vec
    vec = data[data.index == target_url]
    # determine similarity
    if mode == "scalar":
        data["sim"] = data.apply(lambda row: sum(row.values*vec.values[0]), axis=1)
    elif mode == "cos":
        data["sim"] = data.apply(lambda row: sum(row.values*vec.values[0])/
            sqrt(sum(row.values**2))*sqrt(sum(vec.values[0]**2)), axis=1)
    else:
        print("unknown mode")
        return None
    # sort by similarity
    data = data.sort_values("sim", ascending=False)
    # return top n urls
    return list(data.index)[1:n+1]
```

3.5 similarity prüfen

- input:
https://grapplinginsider.com/john-danaher-picks-the-best-open-guard/
- output:
https://grapplinginsider.com/gordon-ryan-explains-how-to-improve-your-butterfly-and-open-guard/,
https://grapplinginsider.com/wno-ryan-vs-diniz-full-card-play-by-play/,
https://grapplinginsider.com/john-danaher-details-his-systematic-approach-to-no-gi-guard-passing/,
https://grapplinginsider.com/gordon-ryans-half-guard-instructional-dropping-soon/,
https://grapplinginsider.com/polaris-11-results-play-by-play/,
https://grapplinginsider.com/third-coast-grappling-kumite-vi-review-hugo-wins/,
https://grapplinginsider.com/learn-the-seven-most-simple-guard-passes/,
https://grapplinginsider.com/xande-ribeiro-explains-the-side-closed-guard/,
https://grapplinginsider.com/combat-jiu-jitsu-worlds-the-middleweights-results/,
https://grapplinginsider.com/no-gi-worlds-2019-results-adam-wardzinski-sets-biggest-points-differ

⇒ John Danaher ist der Trainer von Gordon Ryan. Die ähnlichen Artikel handeln häufig über Technik Informationen die diese beiden geben. Es sind auch einige Artikel über Wettkämpfe dabei.

- input:
https://grapplinginsider.com/ryan-hall-still-cant-get-anyone-to-fight-him/
- output:
https://grapplinginsider.com/breaking-conor-mcgregor-v-donald-cowboy-cerrone-official-for-ufc-246/
https://grapplinginsider.com/watch-every-kneebar-finish-in-ufc-history/
https://grapplinginsider.com/khabib-nurmagomedov-v-tony-ferguson-is-on/
https://grapplinginsider.com/video-bj-penn-involved-in-second-bar-brawl-this-year/
https://grapplinginsider.com/rener-gracie-recaps-logan-paul-vs-floyd-mayweather-from-jiu-jitsu-pe
https://grapplinginsider.com/demian-maia-v-gilbert-burns-in-the-works/
https://grapplinginsider.com/nick-diaz-announces-return-to-mma/

https://grapplinginsider.com/georges-st-pierre-adcc-anderson-silva/
https://grapplinginsider.com/demian-maia-not-ready-to-retire-yet-plans-to-compete-in-jiu-jitsu/
https://grapplinginsider.com/yamauchi-sets-new-record-at-bellator-229/

⇒ Ryan Hall ist ein UFC Wettkämpfer. Alle ähnlichen Artikel handeln von UFC Wettkämpfen, allerdings nicht von seinen.

- input:
https://grapplinginsider.com/5-reasons-why-marcelo-garcia-is-the-greatest-of-all-time/
- output:
https://grapplinginsider.com/euros-2020-mikey-musumeci-enters-the-openweight-division/
https://grapplinginsider.com/making-the-case-5-reasons-why-roger-gracie-is-the-greatest-of-all-ti
https://grapplinginsider.com/marcus-buchecha-almeida-gle-and-leo-vieira-get-us-citizenship/
https://grapplinginsider.com/exclusive-elisabeth-clay-constant-competitor-won-no-gi-pans-on-likel
https://grapplinginsider.com/mo-jassim-adcc-2019-seth-daniels/
https://grapplinginsider.com/wno-championships-preview-will-mikey-musumecis-dominance-continue-at
https://grapplinginsider.com/brianna-ste-marie-we-deserve-just-as-much-of-a-platform-as-the-men/
https://grapplinginsider.com/euros-2020-3-things-to-look-out-for-on-the-final-day/
https://grapplinginsider.com/watch-roger-gracie-v-andre-galvao/
https://grapplinginsider.com/tom-deblass-retirement-reddit-adcc/

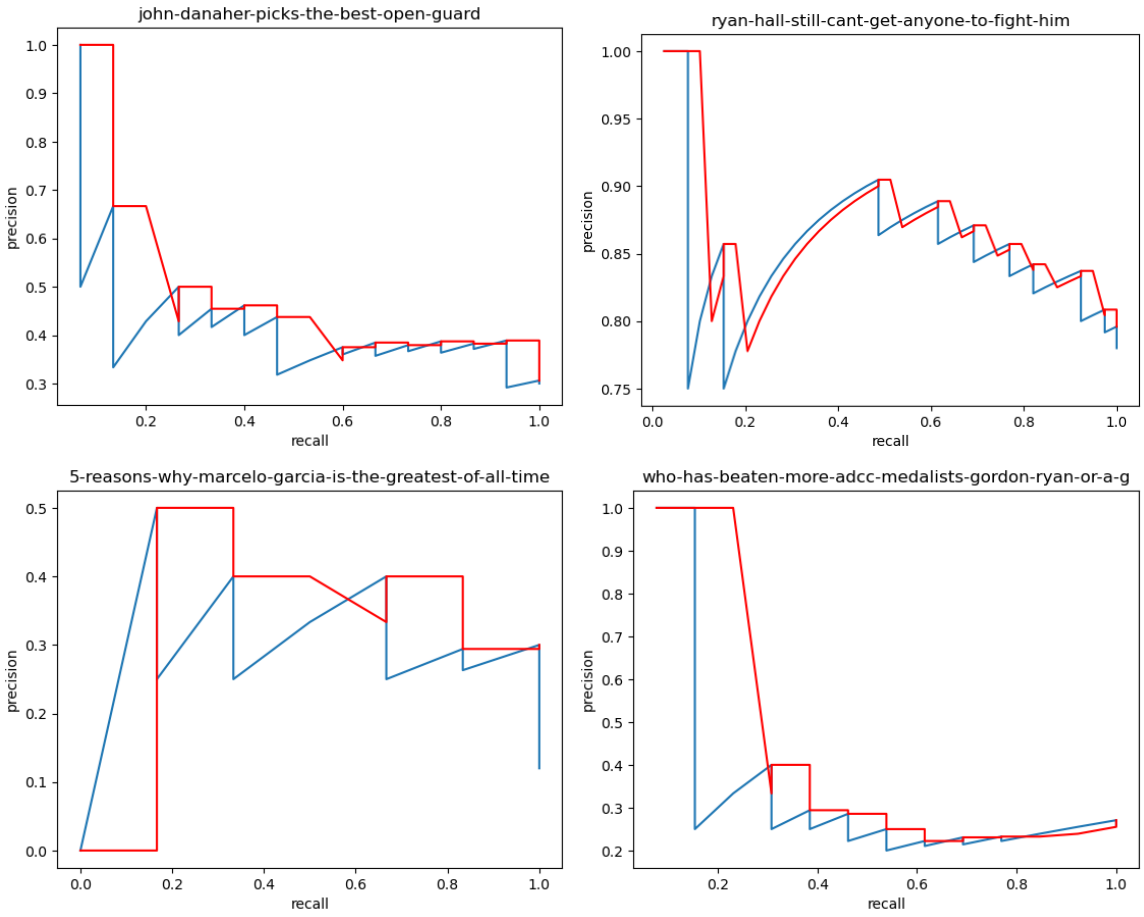
⇒ Marcelo Garcia ist ein BJJ Wettkämpfer. Es geht um einige andere Wettkämpfer die ebenfalls Veteranen sind und heute im Ruhestand. Es sind auch ein paar aktuelle Wettkämpfe dabei, die ich dem ursprünglichen Artikel nicht zuordnen kann.

- input:
https://grapplinginsider.com/who-has-beaten-more-adcc-medalists-gordon-ryan-or-andre-galvao/
- output:
https://grapplinginsider.com/year-end-awards-who-was-the-best-male-no-gi-grappler-in-2021/
https://grapplinginsider.com/year-end-awards-who-was-the-best-female-no-gi-grappler-in-2021/
https://grapplinginsider.com/meet-the-stacked-bjj-stars-v-heavyweight-grand-prix-line-up/
https://grapplinginsider.com/2021-ibjjf-no-gi-worlds-preview-the-loaded-middleweight-division/
https://grapplinginsider.com/demian-maia-to-face-alex-cowboy-oliveira-in-first-grappling-match-si
https://grapplinginsider.com/bjj-stars-8-previewing-the-loaded-8-man-gi-grand-prix/
https://grapplinginsider.com/2021-ibjjf-pans-preview-the-loaded-featherweight-division/
https://grapplinginsider.com/jessa-khan-to-make-one-championship-debut-against-amanda-tubby-alequ
https://grapplinginsider.com/diego-pato-oliveira-leaves-cicero-costha-joins-dream-art-project/
https://grapplinginsider.com/andre-galvao-inducted-into-the-adcc-hall-of-fame/

⇒ ADCC sind quasi die Olympischen Spiele des No-Gi Submission Grapplings. Alle ähnlichen Artikel handeln ebenfalls von No-Gi Submission Grappling. Es schleichen sich kleine Ausnahmen wie ein UFC Match ein.

3 Praktikum 3 - Teil II: Evaluation

3.1 similarity precision/recall



Die blaue Linie ist precision@k und recall@k. Die rote Linie ist die beste precision je recall Wert. Bei dieser Auswertung wurde die gegebene Kategorie `BJJ News` ignoriert, da fast jeder Artikel dieser Kategorie teil ist und somit nahezu ausschließlich true positives zurück kämen. Da unsere Kategorien softe Kategorien sind, sprich sich nicht gegenseitig ausschließen, ist diese Auswertung leider nicht so aussagekräftig wie sie sein könnte.

3.2 Interpretation

Die Qualität der zurückgegebenen Dokumente tendiert dazu mit steigendem k abzunehmen, ist jedoch aufgrund der schlechten Kategorien Labels der Webseite so nicht sauber zu evaluieren.

3 Praktikum 3 - Teil III: Duplikaterkennung mit Shingling und MinHashing

3.1 Schritte nachvollziehen

Aus jedem Dokument wird ein Set von Shingles erstellt und gehasht. Es werden zuerst die Jaccard-Ähnlichkeiten berechnet, dann werden sie mit dem MinHash Algorithmus angewandt.

3.2 Verbesserungen

- Im vorgegebenen Programm ist die Anzahl der Hash-Funktionen (und dementsprechend die Anzahl der Signaturen) zu klein: Wenn nur 2 Signaturen benutzt werden, dann gibt es nur drei mögliche Werte für EstJ (estimated Jaccard similarity): 1, 0.5, 0. Wenn man stattdessen 3 Signaturen benutzt, dann bekommt man 4 mögliche Werte für EstJ: 1, 0.67, 0.33 und 0.
- Zusätzlich kann man den Threshold erhöhen, dann entsteht aber die Gefahr, einige True Positives zu verlieren.
- Außerdem kann man die Fenstergröße (i.e., die Anzahl der Wörter in einem Shingle) variieren.

vorher:
`numHashes`=2, `window_length`=3, `threshold`=0.5
⇒ `true_positive`=10, `false_positive`=5, `true_negative`=985, `false_negative=0`
⇒ `precision`=$\frac{2}{3}$, `recall`=1

nachher:
`numHashes`=4, `window_length`=3, `threshold`=0.76
⇒ `true_positive`=10, `false_positive`=0, `true_negative`=990, `false_negative`=0
⇒ `precision`=1, `recall`=1

3.3 Zeichenbasiert statt Wortbasiert

zeichenbasiert:
`numHashes`=4, `window_length`=20, `threshold`=0.5
⇒ `true_positive`=10, `false_positive`=0, `true_negative`=990, `false_negative=0`
⇒ `precision`=1, `recall`=1

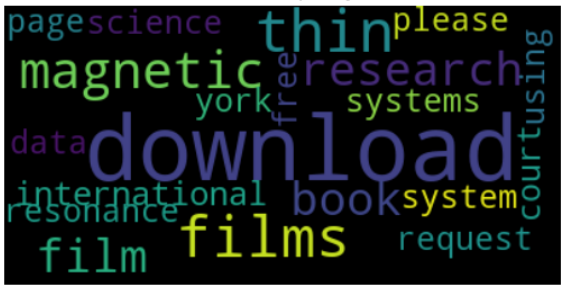
Bei dem zeichenbasierten Shingling dauert die Berechnung der wirklichen Jaccard-Ähnlichkeit sehr lange - 113.3 Sekunden vs. 5.5 Sekunden für das approximative Vorgehen mit den Signaturen (für 1000 Dokumente). Das hat mit der vergrößerten Anzahl von Shingles in jedem Dokument zu tun - durchschnittlich 1560.67 Shingles pro Dokument, wobei es bei dem wortbasierten Shingling nur 251.24 Shingles pro Dokument gab. Wenn man eine deutlich kleinere Fenstergröße nimmt (z.B. 6 Zeichen), dann bekommt man nur eine wenig kleinere Anzahl an Shingles pro Dokument (1432.5), das hat aber - wie erwartet - eine schlechte Auswirkung auf das Ergebnis (30 False Positives).

4 Praktikum 4 - Teil I: Topic Modell Parameter und Interpretation (Common-Crawl)

4.1 Topic Überbegriffe

0: medical research

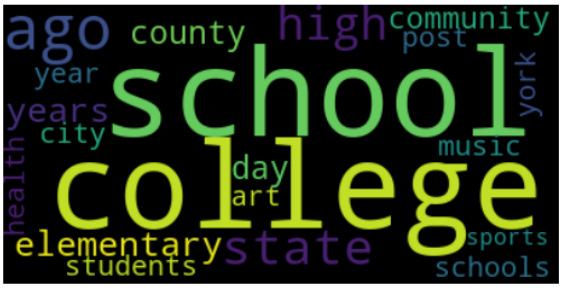
1: media piracy



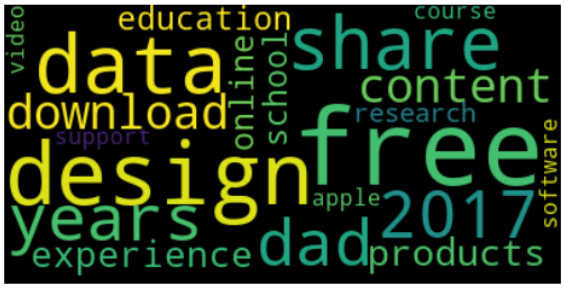
2: blogging

3: self help / DIY

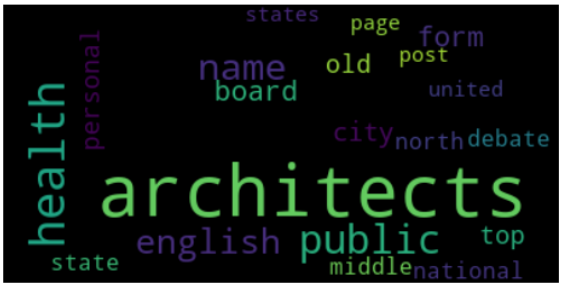
4: education (CORRECTION: college sports teams)



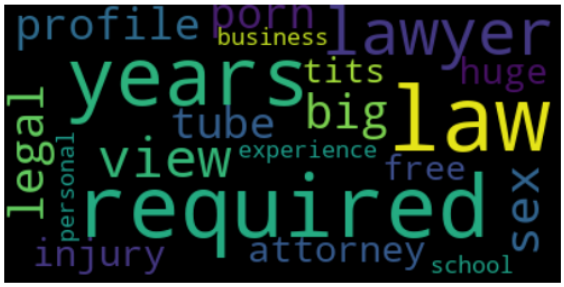
5: design

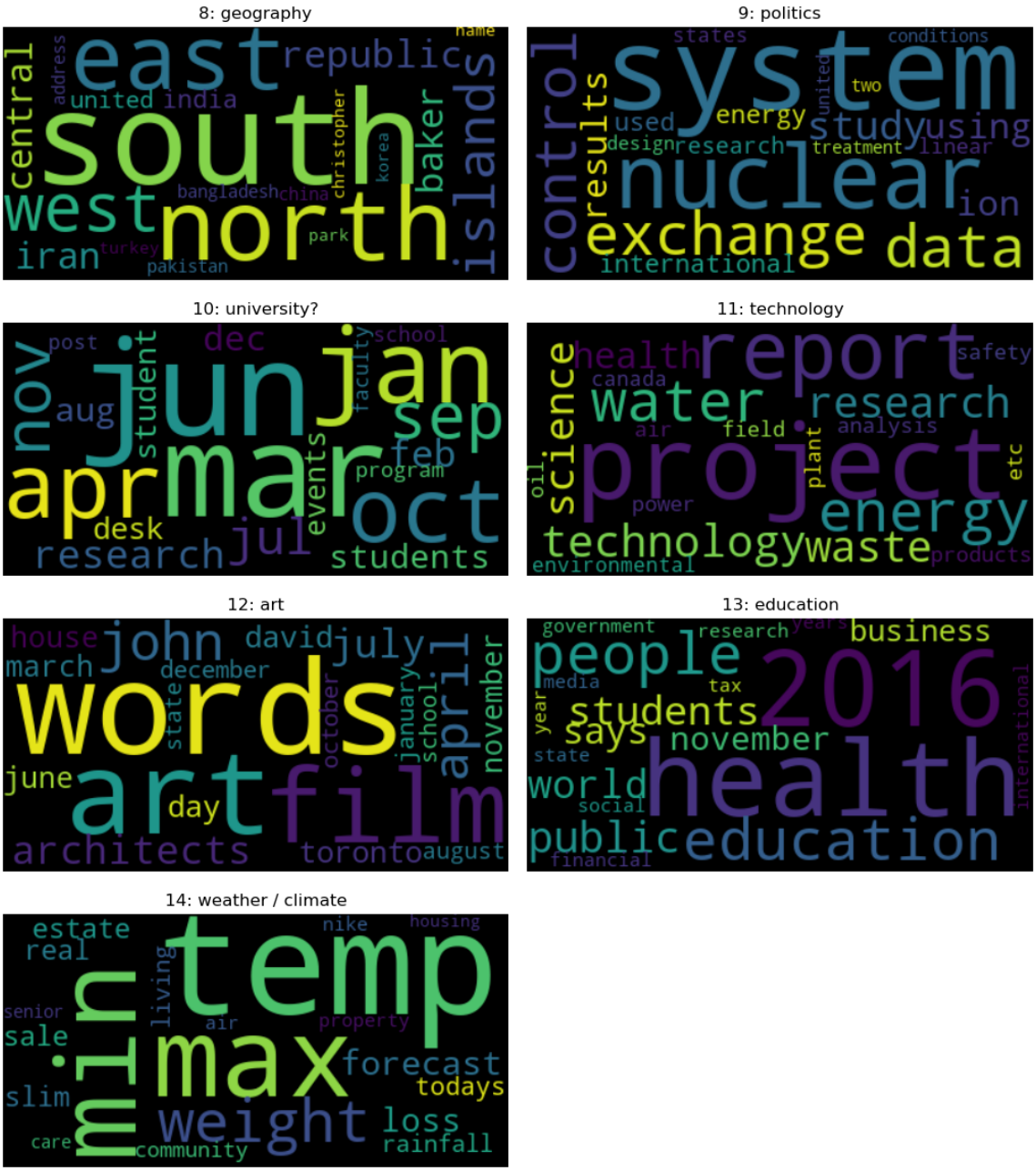


6: architecture



7: adult





Mit Ausnahme von 10: University erscheinen alle relativ sinnvoll, unter Berücksichtigung dass sich die Qualität der Topics unterscheidet

4.2 Filter Adult Topic

```
censored_result = result[(dfnormal[7] <= 0.5) | (dfnormal[7].isna())]
```

```
censored_dfnormal = dfnormal[(dfnormal[7] <= 0.5) | (dfnormal[7].isna())]
```

Wir entfernen hier aus dem CC Datensatz alle Dokumente mit einer Wahrscheinlichkeit über 50% des Adult Topics (7).

4.3 Stichproben

Education Topic (4) mit über 70% Wahrscheinlichkeit:

http://lmmc.ca/en/concert_details.php?concert_id=114
https://www.620ckrm.com/2017/03/11/regina-cougars-wbb-team-to-play-in-usports-canada-consolation-
http://feathermerchant.com/?category=sports%3Ecollege&id=363&type=golf%20towel
https://www.thestudentroom.co.uk/showthread.php?t=5012052
http://ginasbluewaterbabies.com.au/category/uncategorized/page/2/
http://gamestrailer.info/universities-in-newcastle-upon-tyne/
https://www.delgazette.com/wire/state-wire/55244/ohio-student-athlete-badly-hurt-in-makeshift-pool
http://osuvetjobs.org/jobs/10946747/experienced-vet-tech-needed-in-busy-4-doctor-practice
https://www.roero-illuminazione.it/cms/community/social/immagini/foto-community.html
http://www.kofc.org/en/columbia/detail/2012_08_legacy.html

Wir würden hier den Topic Überbegriff von Education Topic (4) auf College Sports Teams Topic (4) korrigieren. Wir haben uns auch einige andere angeschaut, mussten aber sonst keine Korrigierungen vornehmen.

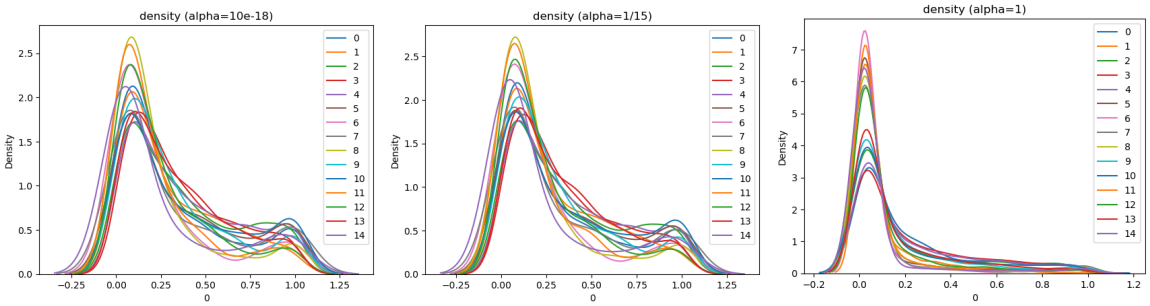
4.4 Topic Mischungen

Blogging Topic (2) und College Sports Teams Topic (4) zu jeweils mindestens 40%:

https://nevadacycling.wordpress.com/2012/09/
https://navarrepress.com/tag/races/
http://thepipelineshow.blogspot.com/2012/04/whl-dominates-chl-attendance-numbers.html
http://southfloridasport.blogspot.com/2016/02/miami-ratings-january-2016.html
https://postinspostcards.com/2015/12/02/cfa-2015-game-103-cincinnati-at-usf-nov-20/
http://photos.gardner-webb.edu/2016-Photos/December-2016/Shanghai-Faculty-Visit/i-gKr3BXd/
http://unitykhartoum.blogspot.com/2014/09/an-exchange-of-emails-between-dr-marina.html
https://buifc.wordpress.com/upcoming-events/
http://eagles-rju.blogspot.com/2018/06/four-ncaa-skaters-among-first-14-taken.html
http://eagles-rju.blogspot.com/2015/09/

Die erhaltenen Dokumente sind Sports Blogs, somit exakt was wir gesucht haben.

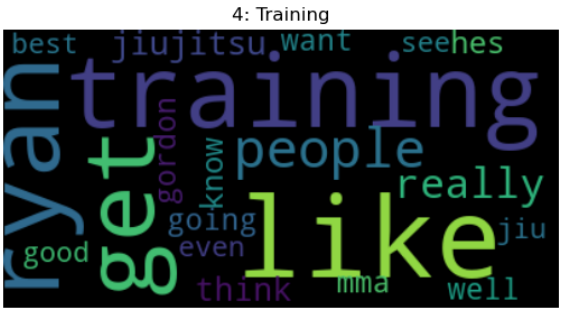
4.5 Zwei neue Modelle



Geplottet sind hier für die drei Modelle je Topic die Häufigkeiten für Wahrscheinlichkeitswerte im Corpus. Wie man sieht verschwindet mit zunehmendem Glättungsparameter der zweite Höhepunkt um die 1. Das liegt daran, dass pro Dokument die Topicwahrscheinlichkeiten geglättet werden, das Modell somit weniger annimmt, dass dem Dokument nur ein Topic unterliegt.

4 Praktikum 4 - Teil II: Topic Modell auf eigenen gecrawlten Texten

4.1 Topic Überbegriffe



Erscheinen alle sinnvoll, wobei wir hier auch eine Topicanzahl gewählt haben über die es hinaus weniger sinnvoll wurde.

4.2 Filter Adult Topic

Unser Datensatz beinhaltet keinen Adult Content.

4.3 Stichproben

Da wir in unserem Datensatz die Titel haben, haben wir diese statt den Links ausgegeben:

```
0 MMA
Video: On closer look, Aljamain Sterling appears to tap to Damien Nitkin at High..
Who's Next episode 6 results and recap: The final is set
'Ready to get famous?': Whos Next episode 1 recap and results
Mike Tyson, Wiz Khalifa, and Nate Diaz sponsor athletes at High Rollerz 15
FloGrappling's Whos Next: Submission Fighter Challenge reality show set for May..
Chael Sonnen Hints That Submission Underground Will Move to FloGrappling
Craig Jones Vs. UFC's Donald 'Cowboy' Cerrone In Combat Jiu-Jitsu Worlds 2021
Dillon Danis Ejected From UFC 268, Slapped by Manager Ali Abdelaziz
High Rollerz: Cops Vs Stonerz      18 September!
Demian Maia Not Ready to Retire Yet, Plans to Compete in Jiu-Jitsu
```

⇒ Passt in etwa, aber sehr hit and miss.

```
1 Streaming
Gordon Ryan insists on no-time-limit stipulation in fourth Felipe Pena match
2022 IBJJF no-gi Pans recap and black belt results
Weekend grappling recap: Polaris 21 and IBJJF Atlanta Open results
Renato Canuto vs. Tommy Langaker booked for ONE 160
WNO: Gordon Ryan vs. Felipe Pena full event results and video highlights
FloGrappling releases statement about Ryan vs. Pena match controversy
Sub Only Series VII results and highlights: PJ Barch submits four in a row
Josh Cisneros vs. Damien Anderson booked for Fight To Win title
WNO: Gordon Ryan vs. Felipe Pena full card line-up and preview
Dates and location set for 2022 IBJJF no-gi Worlds
```

⇒ Passt in etwa auf Event Streaming.

```
2 BJJ
Polaris 20: USA Vs Brazil Results
UFC 274: Oliveira Submits Gaethje with RNC
Polaris 19 results and highlights: Roberto Jimenez dominates, Ash Williams and..
Emerald City Invitational results and highlights: Underdog Kieran Kichuk wins $10K
SUG 29 Results: Andy Varela Taps Sean Strickland in Bizarre Main Event
Polaris 18 Results and Video Highlights: Ashley Williams Upsets Paulo Miyao
Raw Grappling 1 Results and Highlights: Yuri Simoes Wins 8-Man Grand Prix
SUG 28 Results and Highlights: Andy Varela Upsets Haisam Rida in Main Event
Polaris 17 Results and Highlights: Craig Jones Retains Title Against Cautious Davi..
SUG 27 Results and Video Highlights: Mason Fowler Retains Title Against Gabriel..
```

⇒ Sollte eher Ergebnisse und Highlights heißen.

```
3 NO-GI
Mikey Musumeci beats Cleber Sousa to capture first-ever ONE submission grappling..
Nick Rodriguez, Giancarlo Bodoni and others set for a stacked EBI 20
Match preview: Mikey Musumeci vs. Cleber Sousa for first-ever ONE submission..
Full main card revealed for WNO: Pedro Marinho vs. Giancarlo Bodoni
Mikey Musumeci vs. Cleber Sousa booked for first ever ONE submission grappling..
Gordon Ryan Vs Felipe Pena: WNO Announce Major Line Up
Jacob Couch vs. Eoghan O'Flanagan set for Grapplefest championship
Giancarlo Bodoni replaces Andy Varela to face Jacob Rodriguez at Who's Number One
Jessa Khan to make ONE Championship debut against Amanda 'Tubby' Alequin
Tim Spriggs stripped of WNO title, Gordon Ryan vs. Pedro Marinho now for..
```

⇒ Passt sehr gut.

```
4 Training
Like mother, like son: How Rodrigo Marello's mom taught him jiu-jitsu and more
John Danaher Promotes Nathalia Santoro to BJJ Brown Belt
Competition Tips for BJJ Beginners with Emily Eyles
John Danaher, Craig Jones talk about conflict behind DDS split
Kade Ruotolo: 'If I were to roll with Gordon, I don't think he can heel hook me'
ADCC interview: Ash Williams is ready to make history
Mark Zuckerberg Trains Brazilian Jiu-Jitsu
'Buchecha' is confident ONE Championship title shot will 'happen naturally'
Matty Healy Spotted Doing Brazilian Jiu-Jitsu
'I'm on a new level': Renato Canuto is ready for ONE Championship debut against..
```

⇒ Passt sehr gut.

```
5 Gracie Jiu Jitsu / IBJJF
UFC Champ Leon Edwards Awarded BJJ Black Belt
Mikey Musumeci wants grappling match against Demetrious Johnson
Video: Watch Rodrigo Marello's record-setting $50,000 foot lock
Dave Mustaine Promoted to BJJ Brown Belt
Roger Gracie Gets Promoted to 5th Degree Black Belt
Leandro Lo killed in nightclub shooting
Abraham Lincoln: Hall of Fame Wrestler
Ffion Davies becomes the first Brit to ever make black belt final at IBJJF Worlds
```

Owen Livesey Promoted to BJJ Black Belt
Checkmat's Elder Cruz promoted to black belt

⇒ Sollte eher Promotionen und Kontroversen heißen.

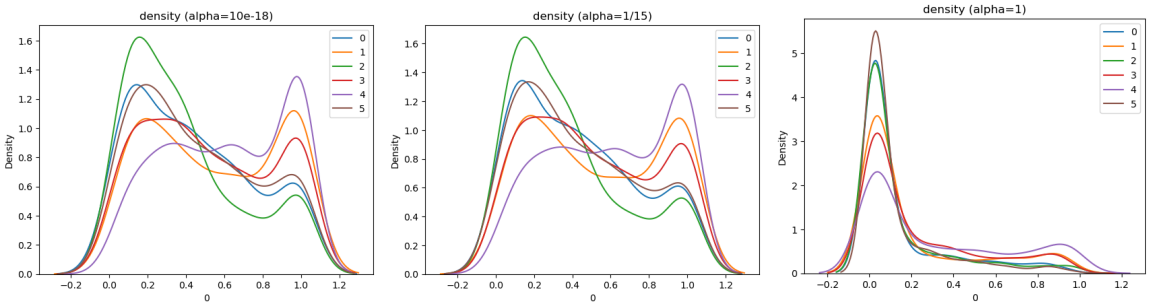
4.4 Topic Mischungen

Results & Highlights Topic (2) und NO-GI Topic (3) zu jeweils mindestens 40%:

Video: Ashley Williams Submits Robert Degle at Grapple Kings 6
Beyond The Match: Re–live Andrew Wiltse's Comeback Submission on Gabriel Almeida
JT Torres, Edwin Najmi to Represent Team USA at Polaris Squads
Polaris Squads to Return With Team USA vs. Team UK and Ireland
Polaris 12: The Undercard's Hidden Gems
ADCC 2019 Results: Craig Jones Chokes Out Thor
ADCC 2019 Results: Craig Jones vs Ben Dyson
Ross Nicholls vs. Vagner Rocha set for Polaris 9

Die erhaltenen Dokumente sind Ergebnisse und Highlights zu No-Gi Wettkämpfen, somit exakt was wir gesucht haben.

4.5 Zwei neue Modelle



Geplottet sind auch hier wieder für die drei Modelle je Topic die Häufigkeiten für Wahrscheinlichkeitswerte im Corpus. Man beobachtet wieder ein mit zunehmendem Glättungsparameter verschwindender zweite Höhepunkt um die 1 aufgrund dessen, dass pro Dokument die Topicwahrscheinlichkeiten geglättet werden und das Modell somit weniger annimmt, dass dem Dokument nur ein Topic unterliegt.