Suchmaschinentechnologie
Semesteraufgabe

Abstract

Die Semesteraufgabe besteht darin eine eigene Suchmaschine mit Solr oder ElasticSearch aufzubauen, die eine Suche innerhalb einer *collection* über COVID-Artikel ermöglichen soll und die anhand von TREC_EVAL evaluiert wird. Die Teilnahme erfolgt in den Teams, die zu Beginn der Veranstaltung erstellt wurden. Als Abschluss der Semesteraufgabe soll (i) eine Ausarbeitung geschrieben und abgegeben werden und (ii) die Ergebnisse im finalen Workshop präsentiert werden.

Das wichtigste auf einen Blick

Die Aufgabe

- Erstellen Sie eine Solr bzw. ElasticSearch Instanz und indexieren Sie die CORD-19 Collection, sodass Anfragen darauf möglich sind
- Nutzen Sie die Topics und die qrel-Daten um mit TREC_EVAL (oder PyTREC-EVAL) die Relevanz Ihrer Retrieval-Ergebnisse zu evaluieren
- Entwickeln Sie eigenständig weitere Verarbeitungsschritte, z.B. Anpassung der Indexierungspipeline, Query-Erweiterungen, weitere (Re-)Ranking-Verfahren, um Ihre Retrieval-Ergebnisse zu verbessern.
- Schreiben Sie eine Ausarbeitung über Ihre Arbeit in diesem Projekt und reichen Sie diese mit ihren 2 besten Runs ein
- Stellen Sie Ihre Arbeit im Rahmen des Workshops am Ende des Semester in einer Präsentation vor

Was ist Wann abzugeben?

- Vorzeigen des MVPs - Minimum Viable Product (Der lange dünne Mann ;-))
  - 09.11.2020
- Abgabe einer schriftliche Ausarbeitung und der finalen Runs
  - 02.02.2021 bis 15:00 Uhr (spätere Einreichungen werden **nicht** akzeptiert)
  - formale Anforderungen der Ausarbeitung
- Halten einer Präsentation
  - 09.02.2021 (Abgabe der Folien nach dem Vortrag)

WICHTIG: Bitte beachten Sie diese Regeln bei der Abgabe Ihrer Resultate

Wie bestehe ich den Kurs

- Einreichung der Ausarbeitung und das halten der Präsentation als Team ist **Pflicht**!
- Abgabe der finalen Runs und Teilnahme an der Challenge als Team ist **Pflicht**!
- Punktevergabe: es gibt insgesamt 100 Punkte zu holen (MVP: 20, Ausarbeitung: 40, Präsi: 40)**.** Bei über 50 Punkten haben Sie bestanden**.**
- Challenge Gewinner und unterstützende Teams erhalten Zusatzpunkte, man kann aber auch 100 Punkte holen ohne gut bei der Challenge abzuschneiden

Zu verwendende Daten und Tools

- Solr bzw. ElasticSearch
- Challenge Daten:
  - CORD-19 Collection (zumindest die *metadata.csv*) zum Indexieren
  - Topics
  - qrel-File
- TREC_EVAL oder PyTREC_EVAL für die Evaluation
- Overleaf (oder LaTeX) für die Ausarbeitung
- PowerPoint, LaTeX, Google Slides, etc. für die Präsentation

Weitere Links:

- Moodle: https://elearning.iws.th-koeln.de/moodle/course/view.php?id=1320
- Teams Tabelle: https://docs.google.com/spreadsheets/d/1xC1YRh8III0guyMyH0R5c54sN6ba3x3e-uenSw5G_0g/edit?usp=sharing

- Solr download: https://lucene.apache.org/solr/downloads.html
- ElasticSearch download: https://www.elastic.co/de/downloads/elasticsearch
- Solr Tutorial: https://lucene.apache.org/solr/guide/8_6/solr-tutorial.html
- ElasticSearch Tutorial: https://www.elastic.co/guide/en/elasticsearch/reference/current/getting-started.html

- CORD-19: https://www.semanticscholar.org/cord19
- TREC_EVAL:
  - download: https://trec.nist.gov/trec_eval/,
  - how to use: http://www.rafaelglater.com/en/post/learn-how-to-use-trec_eval-to-evaluate-your-information-retrieval-system
- PyTREC_EVAL: https://github.com/cvangysel/pytrec_eval
- Anaconda for Python: https://www.anaconda.com/products/individual

- Jupyter Tutorial: https://www.dataquest.io/blog/jupyter-notebook-tutorial/
- Overleaf Tutorial: https://www.overleaf.com/learn/latex/Tutorials

Beschreibung des Projekts

Überblick

Das Projekt beschäftigt sich damit eine Mini-TREC Challenge mit dem CORD-19 (kurz für: COVID-19 Open Research Dataset) abzuhalten.

Ihre Aufgabe ist folgende:

- Nutzen Sie Solr bzw. ElasticSearch, um die *CORD-19* Daten zu indexieren
- Nutzen Sie die *Topics* Daten, um daraus queries zu bauen, die Sie an Ihre Solr bzw. ElasticSearch Instanz abschicken
- Nach jeder Query erhalten Sie eine Resultatliste. Diese Liste können Sie gerne nochmal "re-ranken". Die finalen Runs speichern Sie bitte im TREC_EVAL-konformen Format
- Nutzen Sie die *qrel*-Daten, um mit TREC-EVAL (oder PyTREC_EVAL) zu evaluieren, wie gut Ihre Suchergebnisse sind
- Schreiben Sie eine Ausarbeitung über Ihr Ziel, das Problem, Ihr Vorgehen das Problem zu lösen und Ihre Resultate

Ihr Ziel ist es ein möglichst gutes Suchergebnis zu erreichen sowie die Herangehensweise zu beschreiben und die Ergebnisse sowohl darzustellen als auch zu erklären.

Nutzen Sie hierzu die Informationen aus der Vorlesung, bspw. wie Sie die Daten am besten indexieren, wie sie gute Queries formulieren, oder welche ranking- und re-rankingverfahren Sie verwenden können.

Sie haben das gesamte Semester Zeit dieses Projekt zu bearbeiten.

- Am **09. November 2020** müssen Sie mir Ihr "MVP" kurz vorstellen
- Am **02. Februar 2021 bis 15:00 Uhr** müssen Sie dann Ihre Ausarbeitung und ihre besten "Runs" abgeben.
- Am **09. Februar 2021** findet ein Abschlussworkshop statt, bei dem Sie Ihre Arbeit präsentieren und mit Ihren Kommilitonen sowie mit mir und Herrn Schaer diskutieren...ja, Herr Schaer wird dabei sein ;-)

Bei der Bearbeitung der Semesteraufgabe sind Sie größtenteils auf sich alleine gestellt. Es gibt keine feste Vorgabe, wie Sie dieses Projekt bearbeiten, allerdings sollten Sie sich natürlich nach den Verfahren und Methoden richten, die Sie sowohl in dieser Vorlesung als auch in anderen Veranstaltungen, wie z.B. Information Retrieval, kennengelernt haben. Gerne dürfen Sie auch auf weiteres Fachwissen (z.B. Text Mining), Programme (Python Libraries), Daten (vordefinierte Language Models), Literatur oder Ideen zugreifen – Sie dürfen eigentlich alles machen, außer Schummeln und bestehende Ergebnisse einfach kopieren.

Kommt der **Verdacht des Schummelns** auf, werde ich mich mit dem Team noch vor der Präsentation am 09.02.2021 einen Termin ausmachen und etliche Verständnisfragen stellen.

Bei den weiteren Veranstaltungsterminen am 09.11.2020, 30.11.2020, 14.12.2020, 11.01.2021 biete ich jeweils eine FAQ Runde an, wo Sie Ihre Fragen bezüglich der Semesteraufgabe direkt mit mir besprechen können. Wenn Sie sich aber auch zwischendurch unsicher sind, **stellen Sie Fragen in moodle**, sodass auch alle Teilnehmenden davon profitieren. Bei Fragen persönlicher Natur können Sie mir natürlich zu jeder Zeit eine Email an johann.schaible@th-koeln.de schicken.

Die Daten

CORD-19

CORD-19 ist ein Korpus, sprich eine Collection, akademischer Arbeiten über COVID-19 und die damit verbundene Coronavirus-Forschung. Es wird vom Semantic Scholar-Team am Allen Institute for AI kuratiert und gepflegt, um Text Mining und NLP-Forschung zu unterstützen. Eine weitere ausführlichere Beschreibung finden Sie hier: https://arxiv.org/abs/2004.10706

Die beinhalteten Daten sind:

- **metadata.csv**

Weitere Zusatzdaten, die Sie nutzen können sind:

- full texts
- cord19_specter_embeddings_2020-04-10

Beginnen Sie erstmal nur damit die *metadata.csv* Daten als Ausgangspunkt für die Indexierung in Solr bzw. ElasticSearch zu verwenden. Die weiteren Daten, wie Volltexte und/oder Embeddings können Sie nach Ihrem eigenen Ermessen verwenden. Weitere Informationen zu CORD-19 finden Sie hier: https://github.com/allenai/cord19

Die *Topics* Daten

Die Topics sind quasi der Information Need des Nutzers. Für unsere Aufgabe, gibt es insgesamt 50 Topics mit den folgenden Eigenschaften:

- **topic**: Die Nummer des Topics
- **query**: Die eigentliche Query, wie z.B. “coronavirus origin”
- **question**: Die Frage, die man mit der Query beantwortet haben möchte, z.B. “what is the origin of COVID-19”
- **narrative**: Der Hintergrund warum man diese Frage beantwortet haben möchte. z.B. “Seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans”

Nutzen Sie all diese Information, **aber zumindest die query**, um Anfragen an Ihre Solr bzw. ElasticSearch Instanz zu stellen. Sie können es natürlich händisch machen, eleganter wäre es aber z.B. mit einem Jupyter Notebook, das die Topics Datei automatisiert ausliest und auch die Anfragen automatisiert an Ihre Solr bzw. ElasticSearch Instanz abschickt.

Anhand dieser Topics wird dann auch die Güte Ihrer Suchmaschine bewertet. Mehr hierzu in den nächsten zwei Unterkapiteln zu qrel-Daten und TREC_EVAL.

Relevance Judgements (qrels-Datei)

Diese Datei enthält eine Liste von Dokumenten, die für jedes Topic als relevant betrachtet werden. Das Format eines Relevanzurteils besteht aus folgenden Spalten:

- **topic-id:** ID oder Nummer des Topics aus der Topics Datei
- **iteration** (muss für TREC_EVAL vorhanden sein, hat aber keine Aussage für uns)
- **cord-doc-id:** ID des Dokuments aus der Collection (aus CORD-19 metadata.csv)
- **judgment:** Beurteilung ob das Dokument relevant für das Topic ist.

Bei **judgement** ist das Urteil 0 für nicht relevant, 1 für teilweise relevant und 2 für vollständig relevant. Die **Iteration** erfasst die Runde, in der das Dokument beurteilt wurde. TREC_EVAL verwendet das Iterationsfeld nicht (obwohl es aus historischen Gründen erwartet wird, dass es vorhanden ist). Die qrels-Datei enthält alle bewerteten Dokumente. Dokumente, die nicht in der qrels-Datei enthalten sind (weil sie nicht bewertet wurden), werden automatisch als nicht relevant betrachtet.

TREC_EVAL

TREC_EVAL ist ein Tool zur Auswertung von Rankings, entweder von Dokumenten oder anderen Informationen, die nach Relevanz sortiert sind. Die Auswertung basiert auf zwei Dateien: Die erste Datei ist die ***qrels*** (relevance judgements) und die zweite Datei enthält die Ranglisten der von Ihrem Retrieval Systems zurückgegebenen Dokumente, sprich die **Runs Ihrer Suchmaschine**. Aufgerufen wird TREC_EVAL über die Shell (Unix) oder die Kommandozeile (Windows) mit

- Unix: `./trec_eval qrel_file run_file`
- Windows: `trec_eval.exe qrel_file run_file`

Ausgegeben wird ein Bericht, der die Güte des Retrieval Systems repräsentiert. Einzelheiten hierzu finden Sie in den Vorlesungsmaterialien oder hier: http://www.rafaelglater.com/en/post/learn-how-to-use-trec_eval-to-evaluate-your-information-retrieval-system

TREC_EVAL ist leider etwas aufwendiger in Windows zu installieren. Daher, hier weitere Infos zur Installation in Windows.
https://github.com/usnistgov/trec_eval/blob/master/README.windows.md

Als Alternative können Sie auch **PyTREC_EVAL** nutzen, was sie hier finden können: https://github.com/cvangysel/pytrec_eval

Abgabe der Projektresultate

Bearbeiten Sie Ihr Projekt in den Teams, die zu Anfang der Veranstaltung festgelegt wurden. Insgesamt müssen Sie die folgende vier Arbeitsergebnisse abliefern:

1. Vorzeigen einer ersten lauffähigen Installation und Evaluation, sprich das Minimum Viable Product (MVP), oder auch als “Der lange dünne Mann” bezeichnet, in einer bilateralen Zoom-Session mit mir.
    a. Vorzeigen am 9.11.2020
2. Die **endgültigen Runs Ihrer Suchmaschine**. Hierzu liefern Sie die zwei besten Runs im TREC-Format für die Topics 1-50; insgesamt also 2 Dateien. Im Vergleich zur ersten Evaluation mit dem MVP, wollen Sie hier natürlich ein viel besseres Evaluationsergebnis erreichen. Verwenden Sie hierfür Ihre neu gewonnen Kenntnisse über Indexierungspipelines, Query-Generierung, Rankingverfahren und weiteres.
    a. Abgabetermin: 02.02.2021, 15:00 Uhr
3. Eine **Ausarbeitung**, die einerseits Ihre praktischen Arbeiten dokumentiert und die Ergebnisse und Herangehensweise beschreibt, und andererseits die kritische Analyse Ihrer eigenen Arbeit beinhaltet.
    a. Abgabetermin: 02.02.2021, 15:00 Uhr
4. **Präsentation und Diskussion** Ihrer Arbeit im Workshop am Ende des Semesters
    a. Termin der Präsentation: 09.02.2021

Form der Abgabe

Die Abgabe der endgültigen Runs, der Ausarbeitung und der Präsentation erfolgen über Mail an mich (johann.schaible@th-koeln.de). Für diese drei Abgaben halten Sie bitte folgende Form ein, ansonsten gibt es Punktabzug.

Bitte notieren Sie in Ihrer E-Mail an mich den Teamnamen und die Namen der Teammitglieder!

WICHTIG!!!: Schicken Sie mir eine Übersicht über die erbrachten Leistungen jeder einzelnen Person im Team.

- Also für die Ausarbeitung z.B.:
    - Person A: hat Kapitel 1-3 geschrieben
    - Person B: hat Literatur gesichtet, in Kapitel 4 zusammengefasst und die Literaturliste erstellt.
    - Person C: hat die Einleitung und das Fazit geschrieben und die Endredaktion (Rechtschreib- und Layoutkontrolle) übernommen
- für die endgültigen Runs z.B.:
    - Person A: hat die Indexierungspipelime implementiert
    - Person B: hat das Re-Rankingverfahren entworfen, das das Publikationsjahr berücksichtigt und somit das Retrievalergebnis verbessert
    - Person C: hat die Jupyter Notebooks mit Python Code gefüllt und die Abfragen an Solr abgeschickt
    - Alle: Haben zusammen die Indexierungspipeline konzipiert
- für die Präsentation z.B.:
    - Alle: haben die Folien konzipiert und auf Fehler überprüft
    - Person A: hat Folien 1-5 erstellt
    - Person B: hat Folien 6-10 erstellt
    - Person C: hat Folien 11-15 erstellt

The Minimum Viable Product (MVP)

Das Minimum Viable Product (MVP) beschreibt das einfachste und möglichst schnell genierte Arbeitsergebnis von A bis Z, das sich erstmal nicht mit “Details” rumplagt, aber funktioniert. Für Sie bedeutet dies folgendes:

- Sie installieren Solr bzw. ElasticSearch, Python und TREC_EVAL
- Sie erstellen eine “default” Indexierung der CORD-19 Metadaten (metadata.csv) in Solr bzw. ElasticSearch, sodass diese Daten in der GUI und per API durchsuchbar sind (Nutzen Sie hier schon sehr gerne ein Jupyter Notebook, um über die API Suchergebnisse zu erhalten)
- Erstellung eines ersten *Runs* mit den Queries der Topics und Ausgabe der Suchergebnisse in einem TREC_EVAL konformen Format.
- Eine erste Evaluation des Runs mit TREC_EVAL. Natürlich können Sie auch PyTREC_EVAL verwenden.

Schauen Sie sich die Inhalte der Vorlesung an, was genau ein Run ist und wie man dessen Suchergebnisse in das TREC_EVAL Format transformiert (siehe auch das Example-Notebook in moodle). Selbiges gilt für die Durchführung einer Evaluation mit TREC_EVAL

Was müssen Sie also abgeben?
Ab dem **09.11.2020** mache ich mit jedem Team einen Termin aus, sodass mir jedes Team das MVP vorstellt.

Die endgültigen Runs

Die endgültigen Runs stellen ihr finales Ergebnis dar und wie im Endeffekt die von Ihnen konfigurierte Suchmaschine abschneidet. Im Vergleich zur ersten Evaluation mit der *default Konfiguration* sollten Sie Ihre neu gewonnen Kenntnisse über Indexierungspipelines, Query-Erweiterung, Rankingverfahren und weitere Aspekte der Vorlesung eingesetzt haben.

Hier sollten sie also **mindestens drei** der folgenden Aspekte berücksichtigen:

- Erweiterung bzw. Anpassung der Indexierungspipeline zur besseren Verarbeitung des CORD-19 Datensatzes (bspw. besseres Schema, Stemming, Stop-Words, etc.)
- Analyse der Topics und Erarbeitung einer Strategie zur Extraktion von Fragetermen aus den Topics – maschinelle oder händische Umsetzung der Anfragen ist möglich. Es geht halt auch besser als einfach nur die Titelterme als Query zu nehmen.
- Einsatz von Query Expansions und einer Strategie zur Erzeugung solcher Erweiterungen
- Experimente mit verschiedenen Rankingverfahren und Gewichtungsfaktoren (bspw. Boosting)
- Experimente mit Re-Rankingverfahren. Dies geschieht meist nachdem Solr bzw. ElasticSearch eine erste gerankte Liste zurückgibt. Diese wird dann z.B. mit Python verarbeitet und mit Hilfe von weiteren, “teuren” Methoden neu gerankt.

- Verwendung weiterer Daten und Informationen zum CORD-19 Datensatz, wie z.B. die Volltexte oder die Embeddings
- Ein eigener Ansatz oder eine eigene Idee zur Verbesserung der Retrieval-Ergebnisse. Experimentieren Sie rum :-)

Wann erreiche ich ein gutes Ergebnis?

- **Gut**, wenn Sie besser abschneiden als unsere Baseline, die durch das MVP markiert wird. Die Anfragen sind die unveränderten Queries der Topics, und es wurde die Defaultkonfiguration bezüglich der Indexierung sowie beim Ranking verwendet.
- **Besser**, wenn Sie im Vergleich zu Ihren Kommilitonen besser abschneiden. Wir bilden eine interne Rangliste und küren die Sieger im Rahmen des Workshops am Ende der Veranstaltung.

Was müssen Sie also abgeben?

Die von Ihnen erzeugten zwei besten Runs für die Topics 1-50 speichern Sie bitte als Dateien ab. Reichen Sie diese Dateien bis **zum 02.02.2021, 15:00 Uhr** ein. Spätere Einreichungen werden nicht akzeptiert.Diese Dateien müssen dem folgenden Schema folgen (andernfalls gibt es Punktabzug):

- TEAMNAME-best1.txt
- TEAMNAME-best2.txt

Ausarbeitung

Das Semesterprojekt beinhaltet viele verschiedene Aspekte, die es gut zu dokumentieren gilt. Erstellen Sie hierzu eine Ausarbeitung, die einen Projektbericht, quasi ein Labortagebuch, in einer wissenschaftlich aufbereiteten Form beinhaltet.

Der Projektbericht beinhaltet ihre eigentlichen Arbeiten, die sie gemacht haben (bspw. welche Query-Expansion Verfahren sie angewandt haben) aber auch die Beschreibung, und Dokumentation Ihrer Ergebnisse. Eine wissenschaftliche Aufbereitung Ihrer Arbeit bildet quasi den Rahmen für das was Sie alles gemacht haben und fokussiert sich gleichzeitig auf die kritische Analyse Ihrer eigenen Arbeit.

Leitfragen für die Ausarbeitung sollen u.a. sein:

- Was war das eigentliche Problem, dass Sie in Ihrem Projekt fokussiert haben? Beschreiben Sie mit Ihren eigenen Worten, worum das Projekt geht, welches Problem existiert und wie Sie dieses Problem lösen möchten.
    - (ca. 0,5-1 Seite)
- Wie genau sah Ihre Methode aus, um diesem Problem zu begegnen? Das ist der Hauptteil der Ausarbeitung. Beschreiben Sie hier Ihre Herangehensweise wie Sie die Retrieval-Ergebnisse verbessern wollten. Es muss für den Leser nachvollziehbar sein, warum die Retrieval-Ergebnisse besser (oder auch eben nicht besser) geworden sind. Beschreiben Sie alle Faktoren in Kürze und legen Sie den Fokus auf einen bestimmten Aspekt (bspw. Re-Ranking), der für Ihre Ergebnisse maßgeblich erfolgreich war.
    - (ca. 2-3 Seiten)

- Wie waren Ihre Ergebnisse und Erfahrungen, die Sie im Projekt erreicht haben? Zeigen Sie die Ergebnisse und beschreiben Sie welche Schritte zu diesen Ergebnissen geführt haben.
    - (ca. 1-2 Seiten)
- Woran hat es gelegen, dass manche Ansätze und Ideen erfolgreich und andere weniger erfolgreich waren? Analysieren Sie Ihre Ergebnisse und warum welche Herangehensweisen mal besser und mal schlechter funktionieren.
    - (ca. 1 Seite)
- Wie können Sie dies mit den theoretischen Hintergründen aus Ihrem Studium (z.B. DIS12) in Verbindung bringen? Diskutieren Sie die Verbindung von Theorie, die Sie in Ihren bisherigen Vorlesung gehört haben, und Praxis, die Sie nun hier beim Aufbau Ihrer eigenen Suchmaschine erlernt haben
    - (ca. 0,5-1 Seite).
- Bilden Sie ein Schlusswort (eine *Conclusion*), das Ihre Arbeit kurz und knapp zusammenfasst. Darüber hinaus geben Sie kurz an, was Sie noch hätten machen können, wenn mehr Zeit gewesen wäre (sprich das *Future Work*)
    - (ca. 0,5 Seiten)

Formale Anforderungen der Ausarbeitung

- Die Ausarbeitung darf auf Deutsch oder auf Englisch angefertigt werden
- **6-8 Seiten im LNCS-Format** des Springer-Verlages
- Vorzugsweise in Latex. Die Vorlage hierzu finden Sie direkt bei Springer (http://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines). Nutzen Sie am besten *Overleaf*. Hier können Sie einen kostenlosen Account erstellen und müssen sich nicht mit einer Latex-Installation rumplagen. Ein Tutorial zu Latex mit Overleaf finden sie hier (https://www.overleaf.com/learn/latex/Tutorials).
- Wenn Sie etwas aus externen Quellen beziehen, z.B. eine Idee wie man das Retrievalergebnis verbessert, so ist das total in Ordnung, wenn sie die Quelle auch referenzieren und zwar als volle Zitation im Literaturverzeichnis und nicht als Fußnote.
- Wenn Sie auf externe Literatur verweisen, verwenden Sie einen einheitlichen Zitationsstil (kleiner Tipp: Latex macht das eigentlich automatisch, wenn Sie Bibtex verwenden).

Was müssen Sie also abgeben?
Schicken Sie mir die fertige Ausarbeitung als PDF-Datei per E-Mail bis zum **02.02.2021 15:00 Uhr**. Spätere Einreichungen werden nicht akzeptiert.

Präsentation

Bei der Präsentation im Workshop sollen Sie die Hauptpunkte Ihrer Arbeit präsentieren. Zeigen Sie hier also

- wie Sie Solr bzw. ElasticSearch eingesetzt haben, um die CORD-19 Daten zu indizieren, sprich Ihre *Indexierungspipeline*
- welche Herangehensweisen Sie verwendet haben, um zu versuchen das Retrievalergebnis zu verbessern
- die Ergebnisse Ihrer Runs (z.B. den MAP-Wert)

- warum Ihrer Meinung nach Ihre Herangehensweise zur Verbesserung der Retrieval-Ergebnisse funktioniert (oder auch nicht funktioniert) hat
- was man noch hätte tun können, um die Retrieval-Ergebnisse zu verbessern

Die Präsentation muss von allen Mitgliedern des Teams mit erstellt werden, jedoch muss Sie nicht von jedem der Teammitglieder präsentiert werden. Hier sind Sie vollkommen frei und können selber entscheiden, ob alle im Team, nur ein paar oder nur ein Teammitglied präsentiert.

Für die Erstellung der Präsentation beachten Sie bitte folgende Punkte:

- Nummerieren Sie unbedingt Ihre Folien, damit man sich besser merken kann wo was war
- Nicht zu viel Text/Information pro Folie, um es leicht verständlich für die Zuhörer zu machen. Schreiben Sie bspw. Ihren gesprochenen Text nicht auf. Verwenden Sie lieber Aufzählungszeichen.
- Klare Struktur(!) der Folien. Es muss sich ein roter Faden durch die Folien ziehen, als ob Sie eine Geschichte erzählen.
- Verwenden Sie bevorzugt Bildmaterial, um Ihre Inhalten besser verständlich zu machen (bspw. ein Diagramm, dass Ihre Indexierungspipeline abbildet) und/oder gewisse Punkte zu unterstützen (bspw. ein Balkendiagramm, dass die Evaluationsergebnisse vor und nach einem Re-Ranking zeigt), wann immer dies nützlich ist.
- Anwendungsbeispiele: Komplexe Botschaften lassen sich oft viel einfacher durch ein Beispiel erklären als durch abstrakte Aussagen. Zeigen Sie Ihre Hauptpunkte also eher an einem Beispiel.

<u>Was müssen Sie also abgeben?</u>
Jedes Team muss eine Präsentation vorbereiten und einen **Vortrag am 09.02.2021** halten. Nach dem Vortrag schicken Sie mir bitte die Präsentation per E-Mail mit den Angaben welches Teammitglied welchen Teil der Präsentation generiert hat.

Leicht zu vermeidende Fehler bei der Ausarbeitung und Präsentation

- Rechtschreibung & Grammatik (Nutzen Sie Grammarly)
- Nummerierung der Gliederung konsistent halten
- Fehlende Referenzen in der Liste der Referenzen
- Fehlende Referenzen im Text
- Fehlende Auszeichnung von Tabellen / Abbildungen
- Zeichensetzung und Auszeichnung von Zitaten
- Seitenzahl nicht eingehalten (zu viel / zu wenig)

Punktevergabe

Notwendige Leistungen

Für die erfolgreiche Bearbeitung dieses Semesterprojekts können Sie bis zu **100 Punkte** erlangen. Bestanden haben Sie, wenn (i) alle Abgaben erfüllt sowie die Präsentation gehalten haben und (ii) Sie mindestens **50 Punkte** erreicht haben. Die Benotung sieht folgendermaßen aus:

- 100-95 Punkte: 1,0
- 95-90 Punkte: 1,3
- 90-85 Punkte: 1,7
- 85-80 Punkte: 2,0
- 80-75 Punkte: 2,3
- 75-70 Punkte: 2,7
- 70-65 Punkte: 3,0
- 65-60 Punkte: 3,3
- 60-55 Punkte: 3,7
- 55-50 Punkte: 4,0

Sollte der Eindruck entstehen, dass Sie einfach nur abgeschrieben oder bestehende Lösungen kopiert haben, werde ich Ihr Team einer mündlichen Prüfung unterziehen, bei der Sie Verständnisfragen zu Ihren Lösungen beantworten müssen. Sollten Sie diese Fragen nicht ausreichend beantworten können, erhalten Sie 0 Punkte.

Die Abgabe der finalen Runs ist zwar Pflicht, wird aber nicht benotet. Es kann aber Zusatzpunkte geben, wenn Sie bei der Evaluation sehr gut abschneiden (siehe hierzu das nächste Kapitel über Zusatzpunkte). Sie können aber auch 100 Punkte erreichen, wenn Ihre Evaluationsergebnisse nicht gut sind. Folgende Elemente fließen in Ihre Benotung mit ein:

- Erstellung des MVP (bis zu **20 Punkte**)
    - Aufsetzen von Solr bzw. ElasticSearch (5 Punkte)
    - Default-Indexierung der CORD-19 Metadaten (5)
    - Installieren von TREC_EVAL (5 Punkte)
    - Durchführen einer ersten Evaluation mit TREC_EVAL, der Default-Indizierung und den Queries aus den Topics 1-50 (5 Punkte)
- Die schriftliche Ausarbeitung (**bis zu 40 Punkte**)
    - es existiert ein verständlicher Projektplan bzw. eine Projektskizze, sprich der Rahmen für Ihre Arbeit (max. 5 Punkte)
    - Ihre Herangehensweisen und was sie in Bezug auf die Evaluationsergebnisse bewirken sind klar beschrieben und ggf. illustriert (max. 15 Punkte)
    - die Ergebnisse sind deutlich dargestellt und beschrieben (max. 5 Punkte)
    - die Diskussion der Ergebnisse, sprich warum sie diese Ergebnisse erreicht haben und was man noch hätte tun können, sind klar dargestellt (max. 10 Punkte)
    - Einhaltung der Ausarbeitungsrichtlinien (bspw. nicht zu viele oder zu wenige Seiten) (max. 5 Punkte)
- Die Präsentation (**bis zu 40 Punkte**)
    - verständliche Darstellung Ihrer Herangehensweisen (max. 10 Punkte)

- verständliche Erklärung wie und warum Ihre Herangehensweise die Retrieval-Ergebnisse beeinflusst (max. 10 Punkte)
- Erläuterung was man noch hätte tun können (max. 10 Punkte)
- Einhaltung der Präsentationsrichtlinien (bspw. Nummerierung der Folien) (max. 5 Punkte)
- Aktive Teilnahme am Abschlussworkshop und an der Diskussion aller vorgestellten Arbeiten (max. 5 Punkte).

Zusatzpunkte

Zusätzlich können folgende Punkte positiv in die Bepunktung einfließen. Dies sind Sonderpunkte, die Sie sich verdienen können. Sie können die maximale Punktezahl auch erreichen, mit einem schlechten Retrievalergebnis, solange Sie die obigen Punkte bearbeitet haben und erklären können, warum ein Ansatz nicht erfolgreich war.

Unterstützung untereinander

Zusatzpunkte gibt es hier für die Unterstützung über Teamgrenzen hinweg. Genauer bedeutet dies Hilfe und Unterstützung für andere Gruppen **im Moodle-Forum(!)** anzubieten z.B. wenn Fragen bei technischen Problemen aufkommen oder wenn man auf interessante Tutorials stößt, wie man eine Python-Bibkliothek für ein Re-Ranking nutzen kann. Warum im Moodle-Forum? Damit es **für alle transparent** ist und nicht der Eindruck entsteht, dass alle Gruppen gemeinsame Sache machen. Sobald es intransparent wird, entsteht immer der Verdacht, dass geschummelt wird, und dies sollten Sie natürlich vermeiden. Natürlich dürfen auch nicht komplette Lösungen, wie das selbe Schema zur Indexierung, einfach wiederverwendet werden. Die Hilfe soll eher inspirierende Charakter haben.

Wenn Ihnen ein anderes Team über Moodle geholfen hat (z.B. TREC_EVAL zu installieren), vermerken Sie dies bitte an der richtigen Stelle in Ihrer Ausarbeitung. Dies dient ebenfalls der Transparenz und wird **keine Minuspunkte** nach sich tragen.

Challenge Gewinner

Weitere **Zusatzpunkte gibt es für die Challenge Gewinner**!!! Nur die endgültigen 2 Runs für die Topics 1-50, die Sie rechtzeitig eingereicht haben, werden bewertet. Hierbei gibt es mehr als nur eine Kategorie, also können eventuell viele gewinnen :-)

Im Abschlussworkshop halten wir dann auch eine kleine Siegerehrung ab und küren die Sieger, die sich neben Ehre und Ruhm auch auch einen kleine schönen Überraschungspreis freuen dürfen :-)

Natürlich müssen Ihre Runs zu Ihrer Ausarbeitung passen. Wer hier schummelt, wird disqualifiziert und erhält für den gesamten Kurs 0 Punkte!