**Einleitung**

Die Evaluierung von Informations-Retrieval-Systemen (IR-Systemen) stellt eine große Herausforderung dar, wenn sich die zugrunde liegende Dokumentensammlung kontinuierlich verändert. Klassische Retrieval-Methoden wie BM25 sind zwar bewährt, gehen jedoch oft nicht auf dynamische Veränderungen von Dokumenten und Nutzeranfragen ein. Dies kann dazu führen, dass die Suchergebnisse mit der Zeit an Relevanz verlieren, da sich Inhalte und Nutzerbedürfnisse wandeln.

LongEval ist ein Forschungsprojekt, das sich genau dieser Problematik widmet. Es untersucht, wie IR-Systeme auf Veränderungen in der Dokumenten-Collection und bei den Nutzeranfragen reagieren. Ziel ist es, herauszufinden, wann und wie oft Systeme aktualisiert werden müssen, um dauerhaft gute Suchergebnisse zu liefern, und wie solche zeitlichen Veränderungen in das Ranking einfließen können. Dabei spiegelt LongEval reale, langfristige Veränderungen wider und ermöglicht so eine realistische Bewertung der Systemperformance über verschiedene Zeitpunkte.

**Fragestellung**

Im Zentrum unserer Untersuchung steht die Frage: Wie verhalten sich IR-Systeme, wenn sich die zugrundeliegende Dokumentensammlung und die Nutzeranfragen im Zeitverlauf verändern? Wann und wie sollte ein Suchsystem angepasst oder neu trainiert werden, um weiterhin relevante Ergebnisse zu liefern? Weiterhin wollen wir herausfinden, inwiefern zeitabhängige Merkmale von Suchanfragen genutzt werden können, um das Ranking zu verbessern und besser auf aktuelle Veränderungen zu reagieren.

**Unsere Idee**

Zu Beginn des Projekts hatten wir die Idee, eine Kombination aus klassischen IR-Methoden und moderner KI-Technologie zu nutzen, um die zeitliche Relevanz von Suchanfragen explizit zu berücksichtigen. Insbesondere wollten wir ermitteln, wie stark eine einzelne Suchanfrage von zeitlichen Faktoren abhängig ist — also ob sie sich auf aktuelle Ereignisse, Trends oder zeitgebundene Themen bezieht.

Um dies zu erreichen, haben wir die OpenAI ChatGPT-API verwendet. Für jede Suchanfrage erzeugt diese API einen sogenannten Zeitabhängigkeits-Score zwischen 0 und 1, der angibt, wie „zeitkritisch“ eine Anfrage ist. Dieser Score lässt sich als eine Art Tag verstehen, der die zeitliche Sensitivität einer Anfrage quantifiziert. So können wir Suchanfragen in zeitabhängige und zeitunabhängige Gruppen einteilen und diese Unterscheidung bei der Ranking-Berechnung berücksichtigen.

Unser Ansatz kombiniert diesen Zeit-Tag mit klassischen IR-Scores wie dem BM25-Score und der Häufigkeit der Dokumentenvorkommen in der Collection. Durch eine gewichtete Kombination dieser Faktoren entsteht ein neuer, optimierter Score, der die Relevanz besser widerspiegelt. Die Gewichte für die einzelnen Komponenten lassen sich variabel einstellen und werden im Rahmen einer systematischen Optimierung justiert, um die bestmögliche Ranking-Performance zu erzielen.

Diese Herangehensweise erlaubt es uns, das Ranking dynamisch an die zeitliche Relevanz anzupassen und so die Robustheit und Aktualität der Suchergebnisse zu verbessern.

**Vorgehen**

**Datenzugriff und -aufbereitung**  
Wir greifen auf eine PostgreSQL-Datenbank zu, in der die LongEval-Web-Daten gespeichert sind. Dort laden wir die relevanten Tabellen: Suchanfragen (Topics), Dokumentinformationen und Häufigkeiten der Dokumentenvorkommen. Diese Daten werden auf eine bestimmte Sub-Collection (z.B. März 2023) gefiltert, um eine kohärente Analyse durchzuführen.

**BM25-Ranking als Basis**  
Als Basis verwenden wir BM25, eine bewährte IR-Methode, die jedes Dokument für eine Anfrage mit einem Score bewertet. Diese BM25-Ergebnisse laden wir aus einer komprimierten Datei, um sie mit den Metadaten der Dokumente und Suchanfragen zu verbinden und für die weitere Verarbeitung vorzubereiten.

**Zeitabhängigkeits-Score mit ChatGPT**  
Für jede Suchanfrage verwenden wir die OpenAI ChatGPT-API, um einen Score zu generieren, der die zeitliche Abhängigkeit der Anfrage angibt. Die API gibt eine Zahl zwischen 0 und 1 zurück, die angibt, wie stark die Anfrage auf zeitkritische Inhalte abzielt. Dieser Wert wird als neues Merkmal („Tag“) in unsere Daten integriert.

**Kombination der Scores**  
Wir führen hier eine einfache Pipeline durch, um unsere ursprünglichen BM25-Scores um einen Bonus für zeitlich relevante (also neu erschienene) Dokumente zu erweitern:

1. **Ausreißer eindämmen:** Zunächst begrenzen wir die gemessene Häufigkeit, mit der ein Dokument im gesamten Korpus auftritt, auf einen festen Maximalwert. So verhindern wir, dass einzelne, sehr häufige Dokumente die spätere Berechnung verzerren.
2. **Neuheitswert berechnen:** Aus der (gegebenenfalls gekappten) Häufigkeit leiten wir einen Wert ab, der umso größer ist, je seltener – und damit wahrscheinlicher neuer – das Dokument ist.
3. **Bonus hinzufügen:** Wir addieren diesen Neuheitswert zu unserem ursprünglichen BM25-Score, jedoch nur dann in vollem Umfang, wenn die Suchanfrage selbst sehr zeitkritisch ist (gemessen an einem Zeit-Score zwischen 0 und 1). Bei weniger zeitkritischen Anfragen bleibt der reine BM25-Wert nahezu unverändert.
4. **Zwischenschritte entfernen:** Um Platz zu sparen und den Fokus auf das Wesentliche zu legen, löschen wir die beiden Hilfsspalten für den ursprünglichen Score und den Neuheitswert.
5. **Endgültiges Ranking:** Abschließend sortieren wir alle Ergebnisse innerhalb jeder Anfrage nach dem neuen kombinierten Score, sodass die Dokumente mit dem höchsten zeitbewussten Wert ganz oben stehen.

**Ranking und Ausgabe**  
Die Dokumente werden anhand des kombinierten Scores sortiert und mit einem Rang versehen. Die Ergebnisse werden schließlich in einem TREC-kompatiblen Format exportiert, um eine standardisierte Evaluation zu ermöglichen.

**Evaluation**  
Wir bewerten die Qualität der Rankings mit bekannten Metriken wie nDCG (Normalized Discounted Cumulative Gain), MAP (Mean Average Precision) und Precision@10. Durch die Variation der Gewichtungsparameter identifizieren wir die Kombination, die die höchste Suchqualität erzielt.

**Fazit**

Unser Projekt zeigt, dass eine Kombination aus klassischen IR-Methoden und KI-basierter Analyse der Zeitabhängigkeit von Suchanfragen das Ranking deutlich verbessern kann. Die Nutzung der ChatGPT-API, um für jede Suchanfrage einen Zeit-Tag zu erzeugen, erlaubt eine differenziertere Betrachtung der zeitlichen Relevanz. Die Kombination dieser Informationen mit traditionellen Scores und weiteren Merkmalen führt zu robusteren und aktuelleren Suchergebnissen.

Dieses Vorgehen bietet eine innovative Möglichkeit, dynamische Veränderungen in der Dokumenten-Collection und den Nutzeranfragen systematisch zu berücksichtigen und so die Leistungsfähigkeit von IR-Systemen langfristig zu erhalten