12. Searching & Midterm review

LING-351 Language Technology and LLMs

Instructor: Hakyung Sung

October 7, 2025

*Acknowledgment: These course slides are based on materials by Lelia Glass @ Georgia Tech (Course: Language & Computers)

Table of contents

Searching

Finding a friend

Information need: What happened to my friend *Katie Smith* from elementary school? How to find out?

- Google search? (What queries to use?)

This is one example of information retrieval/search.

Finding a friend

Information need: What happened to my friend *Katie Smith* from elementary school? How to find out?

- Google search? (What queries to use?)
- LinkedIn, Facebook, Instagram?

This is one example of information retrieval/search.

Finding a friend

Information need: What happened to my friend *Katie Smith* from elementary school? How to find out?

- Google search? (What queries to use?)
- LinkedIn, Facebook, Instagram?
- (Probably not super helpful to ask ChatGPT…)

This is one example of information retrieval/search.

Finding a friend

Information need: What happened to my friend *Katie Smith* from elementary school? How to find out?

- Google search? (What queries to use?)
- LinkedIn, Facebook, Instagram?
- (Probably not super helpful to ask ChatGPT…)
- Ask a mutual friend?

This is one example of information retrieval/search.

Intents and information needs

We need to understand that users interacting with technology have various intents.

- Set an alarm for tomorrow at 6.

Intents and information needs

We need to understand that users interacting with technology have various intents.

- Set an alarm for tomorrow at 6.
- Book a flight to Guatemala.

Intents and information needs

We need to understand that users interacting with technology have various intents.

- Set an alarm for tomorrow at 6.
- Book a flight to Guatemala.
- Intent recognition: identify user's intent from what they say

Intents and information needs

We need to understand that users interacting with technology have various intents.

- Set an alarm for tomorrow at 6.
- Book a flight to Guatemala.
- Intent recognition: identify user's intent from what they say
  - closely related to text classification problem

Intents and information needs

Information needs are the user's underlying intentions to find out or learn something. For example, if someone asks:

- What is the capital of Guatemala?
- What's the weather there?
- When did Mandarin and Cantonese diverge?
- What happened to my friend Katie Smith?

all of these are information needs, because the user's goal is to know something

Intents and information needs

When these needs are turned into search queries, they often appear in a shorter or less natural form. For instance, those same needs might become queries like... (your thoughts?)

Intents and information needs

When these needs are turned into search queries, they often appear in a shorter or less natural form. For instance, those same needs might become queries like... (your thoughts?)

- The information need is the question in the user's mind

Intents and information needs

When these needs are turned into search queries, they often appear in a shorter or less natural form. For instance, those same needs might become queries like… (your thoughts?)

- The information need is the question in the user’s mind
- The query is the way they actually type it into a search engine

Evaluating search results

How to quantify success? Some common metrics.

- Precision
- Recall

Evaluating search results

Precision: Percentage of the documents returned that are relevant.

- Search engine gives you 400 documents, 200 of them are actually relevant: precision is 200/400 = 50%.

Evaluating search results

Precision: Percentage of the documents returned that are relevant.

- Search engine gives you 400 documents, 200 of them are actually relevant: precision is 200/400 = 50%.
- Aka positive predictive value (medicine!): what percent of positives are true positives?

Evaluating search results

Recall: Percentage of relevant documents that are returned.

- There are 1000 relevant documents out there; search engine gives 200 of them: precision is 200/1000 = 20%.

Evaluating search results: Summary

$$\textbf{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

$$\textbf{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents in the collection}}$$

Evaluating search results: Summary

$$\textbf{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

$$\textbf{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents in the collection}}$$

Precision $\Rightarrow$ *How accurate the results are*

Recall $\Rightarrow$ *How complete the results are*

Evaluating search results: Trade-offs and measures

- What would happen if a search engine aimed for **100% precision** but didn't care about recall? $\rightarrow$ *It would return very few (or even one) perfectly relevant documents.*

Evaluating search results: Trade-offs and measures

- What would happen if a search engine aimed for **100% precision** but didn't care about recall? $\rightarrow$ *It would return very few (or even one) perfectly relevant documents.*
- What would happen if a search engine aimed for **100% recall** but didn't care about precision? $\rightarrow$ *It would return everything, including all irrelevant documents.*

Evaluating search results: Trade-offs and measures

F-measure: balances precision and recall (because both matter)

$$F_1 = \frac{2 \times P \times R}{P + R}$$

Evaluating search results: Trade-offs and measures

- In reality, search results are also **ranked** — people care most about the precision of the top-ranked results.

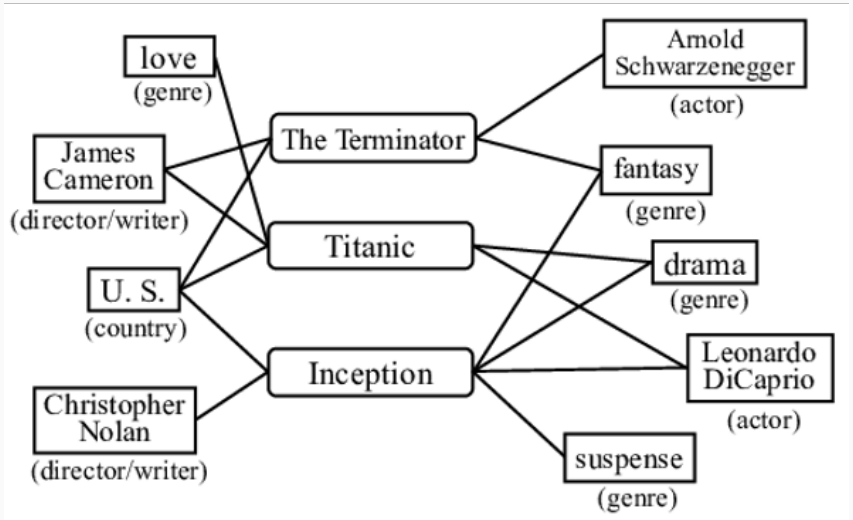
Evaluating search results: Trade-offs and measures

- In reality, search results are also **ranked** — people care most about the precision of the top-ranked results.
- **Precision@k**: looks at the top k results and asks, *what percent of these are relevant?*

Structured data

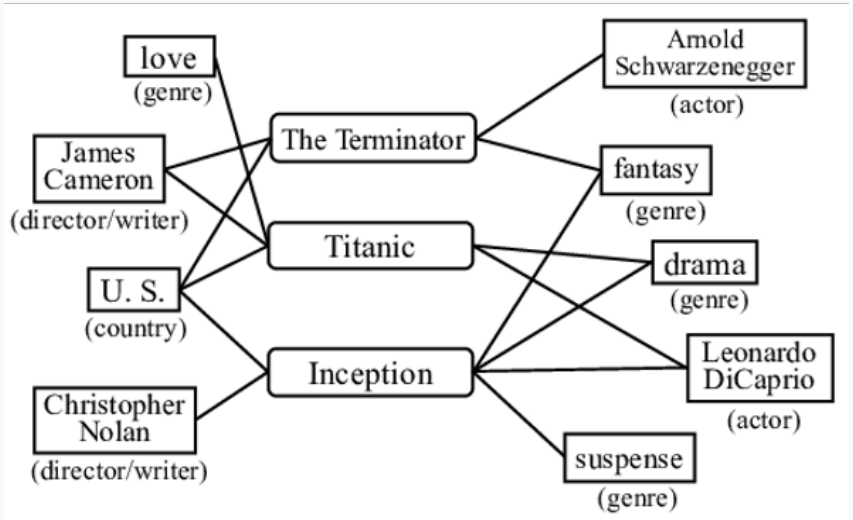
Structured data

- **IMDB**: links each film with year, director, producer, actor, etc.



Structured data

- **IMDB**: links each film with year, director, producer, actor, etc.
- Can visualize in a (limited-domain) knowledge graph (i.e., “ontology”)



Knowledge graphs

- *Google Scholar*: links each paper with year, author, who it cites, who cites it. (`https://scholar.google.com/`)

Knowledge graphs

- *WordNet*: hand-built lexicon representing, e.g., that dogs are animals, animals are living things (hypernym/hyponym relations)

Knowledge graphs

- *WordNet*: hand-built lexicon representing, e.g., that dogs are animals, animals are living things (hypernym/hyponym relations)
- *FrameNet*: a lexical database of English that is both human- and machine-readable (`https://framenet.icsi.berkeley.edu/`)

Knowledge graphs

- *WordNet*: hand-built lexicon representing, e.g., that dogs are animals, animals are living things (hypernym/hyponym relations)
- *FrameNet*: a lexical database of English that is both human- and machine-readable (`https://framenet.icsi.berkeley.edu/`)
- *ConceptNet*: a freely-available semantic network, designed to help computers understand the meanings of words that people use (`https://conceptnet.io/`)

Knowledge graphs

- Google's knowledge graph (`https://www.youtube.com/watch?v=mmQl6VGvX-c`)

Knowledge graphs

- Google's knowledge graph (https://www.youtube.com/watch?v=mmQl6VGvX-c)
- Knowledge graph and searching (https://www.youtube.com/watch?v=Q5izD6Xlb8o)

Midterm

Logistics

- Start Date: October 9, 2025 — 9:00 AM
- End Date: October 9, 2025 — 5:00 PM
- Time Limit: 60 minutes
- Mode: Asynchronous (open within the given window)
- Attempt: 1
- Format: Open-book, but not Open-AI (PLEASE take this as an opportunity to review and apply what you've learned)

Logistics

- Total 19 questions

Logistics

- Total 19 questions
- Each class slide is related to 2-3 questions

Logistics

- Total 19 questions
- Each class slide is related to 2-3 questions
- Multiple choice (10 questions)

Logistics

- Total 19 questions
- Each class slide is related to 2-3 questions
- Multiple choice (10 questions)
- Short written response (8 questions)

Logistics

- Total 19 questions
- Each class slide is related to 2-3 questions
- Multiple choice (10 questions)
- Short written response (8 questions)
- Longer written response (1 question)

Slide guides

- Introduction & Encoding

Slide guides

- Introduction & Encoding
  - What is language

Slide guides

- Introduction & Encoding
  - What is language
  - Different type of languages (alphabetic, syllabic, logographic)

Slide guides

- Introduction & Encoding
  - What is language
  - Different type of languages (alphabetic, syllabic, logographic)
  - language vs. writing (symbols)

Slide guides

- Introduction & Encoding
  - What is language
  - Different type of languages (alphabetic, syllabic, logographic)
  - language vs. writing (symbols)
- Spelling checkers

Slide guides

- Introduction & Encoding
  - What is language
  - Different type of languages (alphabetic, syllabic, logographic)
  - language vs. writing (symbols)
- Spelling checkers
  - Simple checkers

Slide guides

- Introduction & Encoding
  - What is language
  - Different type of languages (alphabetic, syllabic, logographic)
  - language vs. writing (symbols)
- Spelling checkers
  - Simple checkers
  - More complex checkers

Slide guides

- Grammar checkers

Slide guides

- Grammar checkers
  - Syntax

Slide guides

- Grammar checkers
  - Syntax
  - POS

Slide guides

- Grammar checkers
  - Syntax
  - POS
  - Dependency relations

Slide guides

- Grammar checkers
  - Syntax
  - POS
  - Dependency relations
- CALL

Slide guides

- Grammar checkers
  - Syntax
  - POS
  - Dependency relations
- CALL
  - Language learning

Slide guides

- Grammar checkers
  - Syntax
  - POS
  - Dependency relations
- CALL
  - Language learning
  - L1 vs. L2

Slide guides

- Grammar checkers
  - Syntax
  - POS
  - Dependency relations
- CALL
  - Language learning
  - L1 vs. L2
  - transfer

Slide guides

- Grammar checkers
  - Syntax
  - POS
  - Dependency relations
- CALL
  - Language learning
  - L1 vs. L2
  - transfer
  - social factors

Slide guides

- Text as data

Slide guides

- Text as data
    - knowledge-driven

Slide guides

- Text as data
  - knowledge-driven
  - data-driven

Slide guides

- Text as data
  - knowledge-driven
  - data-driven
  - (skip different previous studies)

Slide guides

- Text as data
  - knowledge-driven
  - data-driven
  - (skip different previous studies)
- Corpus, word distributions

Slide guides

- Text as data
  - knowledge-driven
  - data-driven
  - (skip different previous studies)
- Corpus, word distributions
  - (skip English corpora)

Slide guides

- Text as data
  - knowledge-driven
  - data-driven
  - (skip different previous studies)
- Corpus, word distributions
  - (skip English corpora)
  - Word distributions - Zipf, Heap

Slide guides

- Text as data
  - knowledge-driven
  - data-driven
  - (skip different previous studies)
- Corpus, word distributions
  - (skip English corpora)
  - Word distributions - Zipf, Heap
  - Word vectors

Slide guides

- Text classification

Slide guides

- Text classification
  - Different text classification tasks (e.g., spam filtering)

Slide guides

- Text classification
  - Different text classification tasks (e.g., spam filtering)
  - How do we make text classifier?

Slide guides

- Text classification
  - Different text classification tasks (e.g., spam filtering)
  - How do we make text classifier?
  - (skip the details of the perceptron)

Group

Group assigned

- Groups have been assigned and added to myCourses.
- Each group consists of 2–3 members.
- There will be two presentations per group:
  - One member presents in the first round.
  - Another member presents in the second round.
- We will discuss presentation preparation in more detail on **October 16**, after the fall break.
- For now, please stay in touch with your group members and start discussing your ideas.