

HAYSTACK EUROPE 2026 · BERLIN · NETFLIX CINEMATIC EDITION

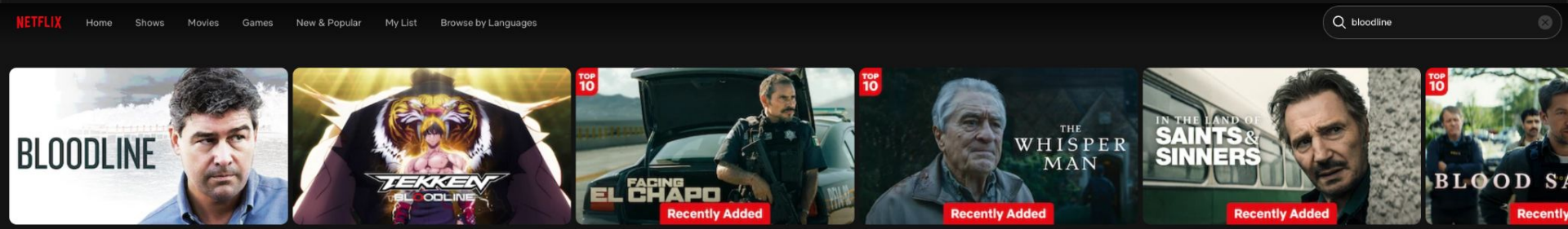
FROM TRIES TO TRANSFORMERS

Scaling Semantic Search

SEASON 1 • EPISODE 01 • DIRECTED BY SEARCH & VOICE TEAM @ NETFLIX • PRESENTED BY IVAN PROVALOV

Starring:
Abhijit Thosar, Anushka Baoni, Claire Campbell, Deniz Turkcapar, Dhinesh Dhanasekaran, Eloise Marszalek, Gaya Varadarajan, Guru Tahasildar, Jeremy Ball, Liyang Hao, Matteo Rinaldi, Noble Kennamer, Shahrzad Naseri, Shreyas Suhas Chavan, Shri Vijayakumar, Spencer L'Heureux, Vito Ostuni

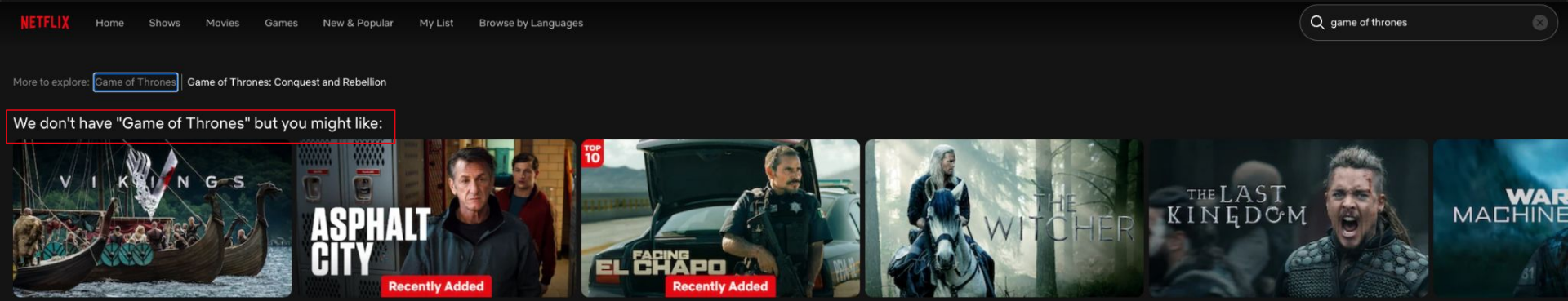
SEARCH AS A SOURCE OF DISCOVERY ON NETFLIX



~16%

OF DISCOVERY STREAMING HAPPENS THROUGH SEARCH

QUERY TRIES & FACETS



Query Facet Mapping (QFM) helps communicate when a show, film or special is unavailable on Netflix when members are searching for it

QUERY FACET MAPPING





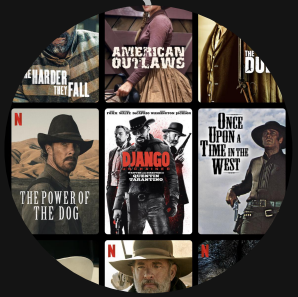
Wes Anderson

person



West Side Story

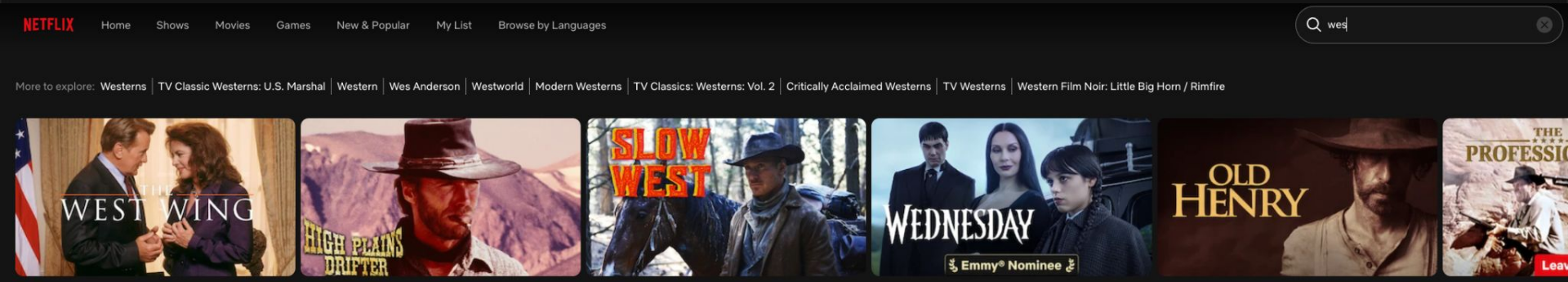
video



Westerns

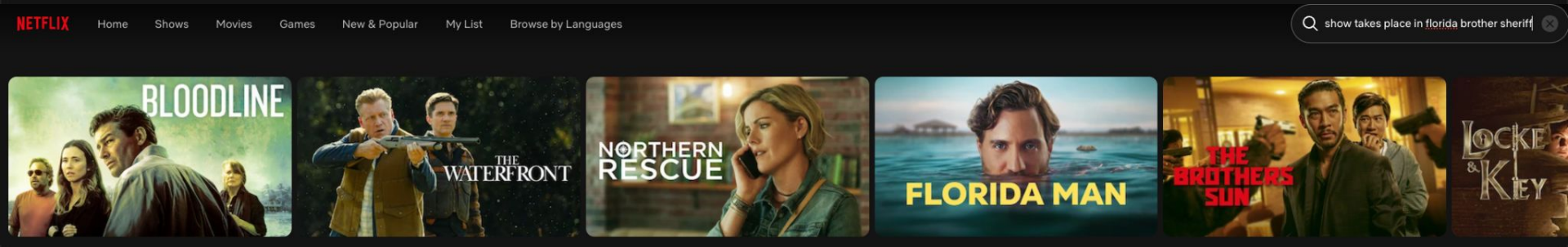
collection

PERSONALIZED SEARCH RECOMMENDATIONS



Ranking for Personalized Search Recommendations is an ML algorithm combining multiple behavioral and lexical ngram-level features

BEYOND NGRAM-BASED SEARCH RECOMMENDATIONS



“netflix show takes place in florida brother sheriff”

THE FOUR DRIVING FORCES

QUERY FORM: KEYWORD

Video, Collection, Person, new intents like a topic

QUERY FORM: NATURAL LANGUAGE

Opens up the intent space significantly to previously unknown intents (e.g.: “add to my list”)

INPUT MODE: VOICE

Allows users to express new intents on a large scale on TV devices previously limited to a click-style keyboard

GUIDED UI JOURNEYS

Chips, pills, allowing user to provide a quick input to the system



TWO CAPABILITIES, WORKING TOGETHER

Serving natural-language search takes both halves of the problem - understanding the query, and retrieving by meaning.

QUERY UNDERSTANDING

Knows what you mean, not just what you type

- Intent & form intent detection
- Slot filling & entity linking
- Spell correction, sensitivity

SEMANTIC RETRIEVAL

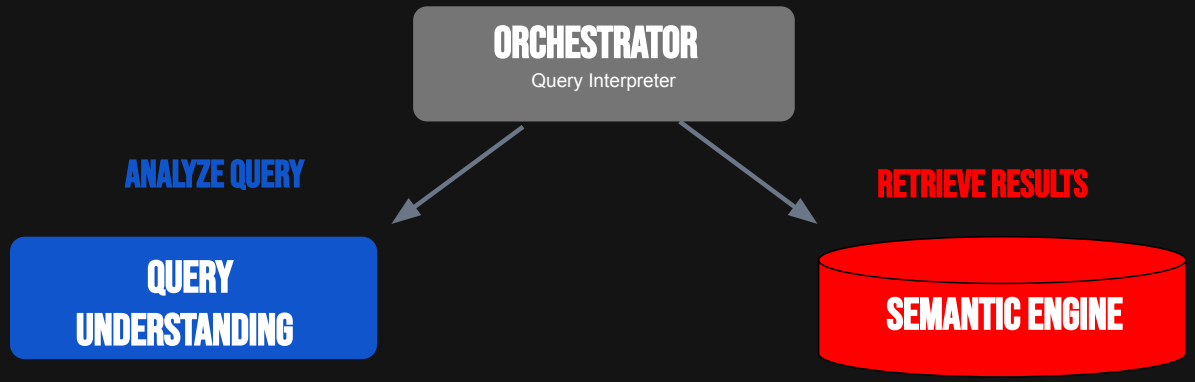
Matches meaning, not exact keywords

- Dense embeddings + lexical, hybrid
- Approximate nearest-neighbor search
- Filtering & boosting to tune precision

SYSTEM ARCHITECTURE

WHERE UNDERSTANDING MEETS RETRIEVAL

We are omitting some details related to ranking tier and many other components of the system.



ROUTING: SPEED MEETS INTELLIGENCE

Queries are routed through three distinct paths depending on complexity and latency requirements.

FAST PATH

Short / simple queries

Trie lookup + entity match

Query Facet Mapping

Returns in under 10 milliseconds

SMART PATH

Complex queries

BERT model

Intents

20-40 ms latency

ADVANCED PATH

Complex / natural language

LLM model: advanced intent, slots

100-1000 ms latency

QUERY UNDERSTANDING

ONE PASS, MANY SIGNALS

A single query goes in ➠ a structured understanding comes out.

QUERY

“Suggest most popular movies to watch with kids”

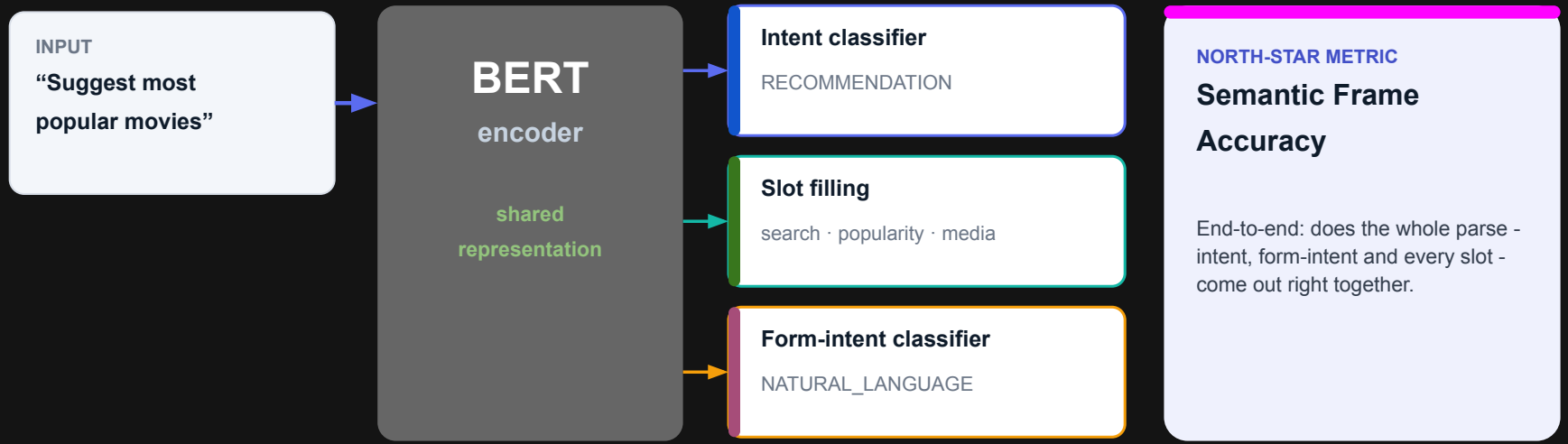


- **Intent detection**
- **Form-intent detection**
- **Entity linking & slot filling**
- **Spell correction**
- **Sensitivity detection**
- **Session classification**
- **Language detection**
- **Query directive**

THE MODEL

ONE BERT ENCODER, THREE HEADS

A multi-task “Joint-BERT” architecture: a compact encoder shared across the tasks query understanding needs.



NORTH-STAR METRIC

Semantic Frame Accuracy

End-to-end: does the whole parse - intent, form-intent and every slot - come out right together.

TAXONOMY

STRUCTURE THE MODEL CAN LEARN

Before a model can parse intent, humans must define it. A rigorous taxonomy of **intents** (the goal) and **slots** (the pieces) is the contract among people, labels and model.

INTENTS: the member's overarching goal

- FETCH
- RECOMMENDATION
- FEEDBACK
- GET_INFO
- SENSITIVE
- UNSUPPORTED

SLOTS: identifiable pieces of the query

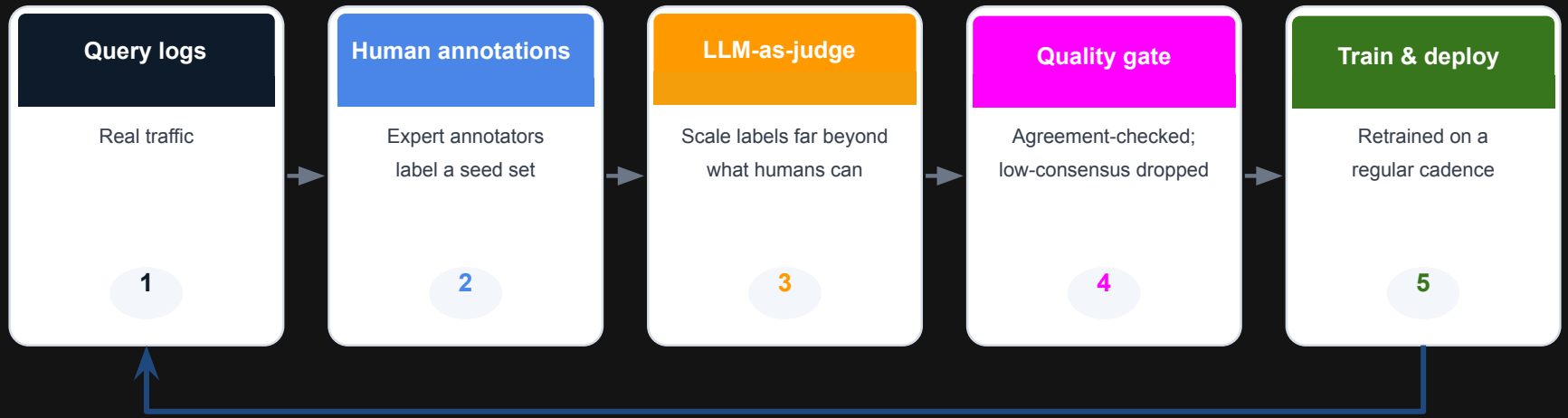
- TITLE
- GENRE
- PERSON
- POPULARITY
- MEDIA
- TOPIC

Living document: the taxonomy keeps evolving — new slots for groups & competitions, and voice intents (navigate, play, transport) next.

FEEDING THE MODEL

THE DATA ENGINE BEHIND THE MODEL

A fully automated pipeline retrains the model on fresh labels continuously.



The unlock: an LLM as a scalable judge lets a small, fast model learn from far more labeled data than human annotation alone could produce.

HUMAN DATA OPERATIONS

TURNING RAW RATINGS INTO GOLD LABELS

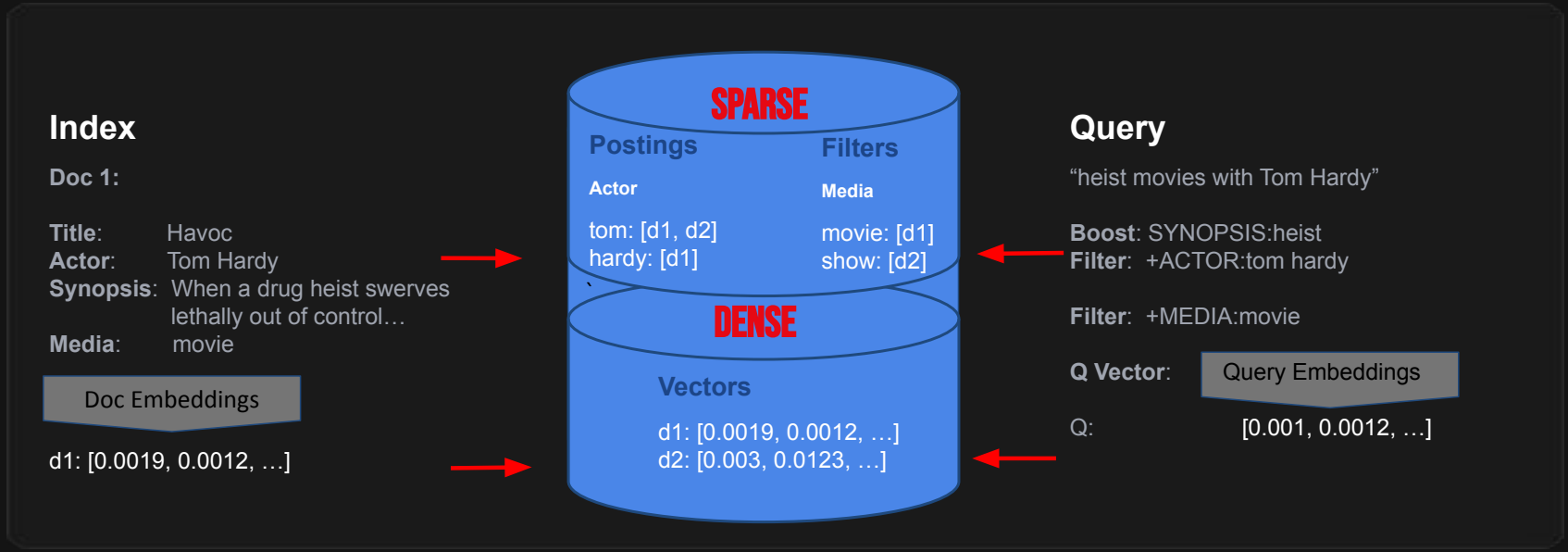
Multiple raters, multiple stages, and a hard consensus gate — quality over quantity.



SEMANTIC RETRIEVAL ENGINE

MATCHING MEANING, NOT KEYWORDS

A hybrid semantic retrieval engine that finds the right titles even when the words don't line up.



IMPACT

WHAT CHANGED FOR MEMBERS

Validated in an A/B test, confirmed with a holdback, then productized across devices.

LONG QUERIES NOW SERVED

Natural-language searches that used to dead-end now return relevant results.

ENGAGEMENT & RETENTION

Statistically significant lift in engagement-mediated retention.

FAST ENOUGH FOR PROD

Understanding runs in the standard-search latency budget, at scale.

A REPEATABLE ENGINE

Taxonomy + annotation + daily retraining keep quality compounding.

TAKEAWAYS

IF YOU TAKE FIVE THINGS HOME

MATCH THE TOOL TO THE QUERY

A trie, a BERT model and an LLM each win somewhere. Route, don't replace.

TAXONOMY IS THE FOUNDATION

A rigorous intent/slot schema is what makes labels and models coherent.

KEEP THE LLM OFF THE HOT PATH

Use it as a judge to scale labels: let a compact model serve traffic. Use at runtime when simpler models won't work.

GUARD LABEL QUALITY RUTHLESSLY

Consensus gates beat raw annotation volume.

HYBRID RETRIEVAL

Dense for meaning, lexical for precision: with filtering and boosting on top.

NEXT EPISODE: Q & A

STREAMING LIVE FROM BERLIN • THANK YOU

iprovalov@netflix.com



CREDITS

QUERY UNDERSTANDING

Query Facet Mapping (QFM)

BERT for Joint Intent Classification and Slot Filling

Modern Information Retrieval, Chapter 5, 7

SEMANTIC RETRIEVAL

Relevant Search

RECOMMENDATIONS

Personalized Search Recommendations @ Netflix