# A Study on Query Expansion with MeSH Terms and Elasticsearch. IMS Unipd At CLEF 2018 eHealth Task 3

**G. M. Di Nunzio**, A. Moldovan

CLEF eHealth Task 3, CLEF 2018, Avignon

AVIGNON CLEF 2018
Conference and Labs for the Evaluation Forum

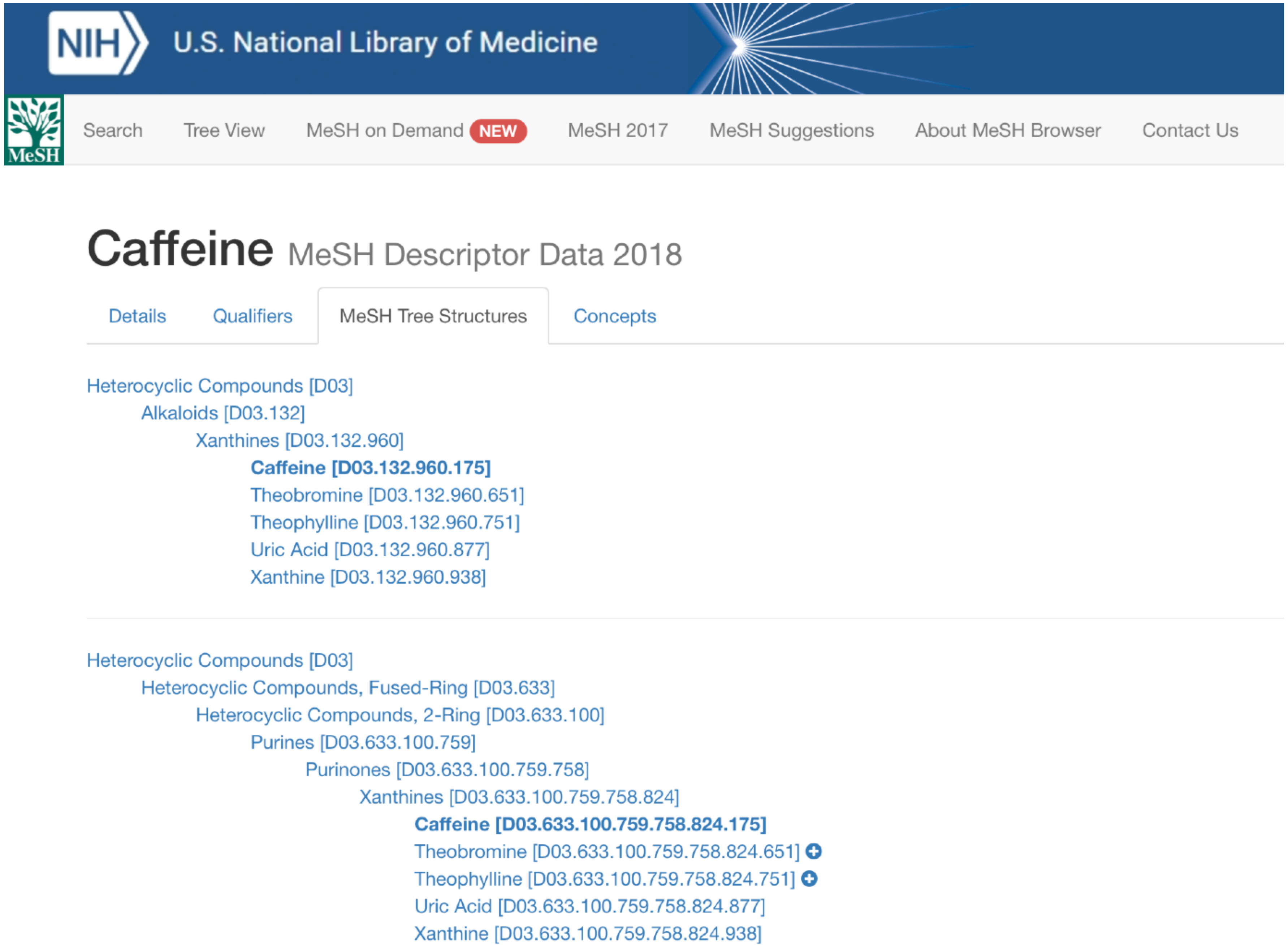
# Objective

- Subtask IRTask 1: Ad-hoc Search
- An evaluation of query expansion approaches that take into account the relationships between MeSH terms
- An evaluation of different document scoring strategies given the multiple ranking list produced by the query expansions

# Query expansion

- Identification of MeSH terms in a query
  - MeshOnDemand
- For Topic 188001 ``caffeine high blood pressure''
  - Caffeine
  - Hypertension
  - Blood pressure

# Query expansion



## Caffeine MeSH Descriptor Data 2018

Heterocyclic Compounds [D03]
  Alkaloids [D03.132]
    Xanthines [D03.132.960]
      **Caffeine [D03.132.960.175]**
      Theobromine [D03.132.960.651]
      Theophylline [D03.132.960.751]
      Uric Acid [D03.132.960.877]
      Xanthine [D03.132.960.938]

Heterocyclic Compounds [D03]
  Heterocyclic Compounds, Fused-Ring [D03.633]
    Heterocyclic Compounds, 2-Ring [D03.633.100]
      Purines [D03.633.100.759]
        Purinones [D03.633.100.759.758]
          Xanthines [D03.633.100.759.758.824]
            **Caffeine [D03.633.100.759.758.824.175]**
            Theobromine [D03.633.100.759.758.824.651] ⊕
            Theophylline [D03.633.100.759.758.824.751] ⊕
            Uric Acid [D03.633.100.759.758.824.877]
            Xanthine [D03.633.100.759.758.824.938]

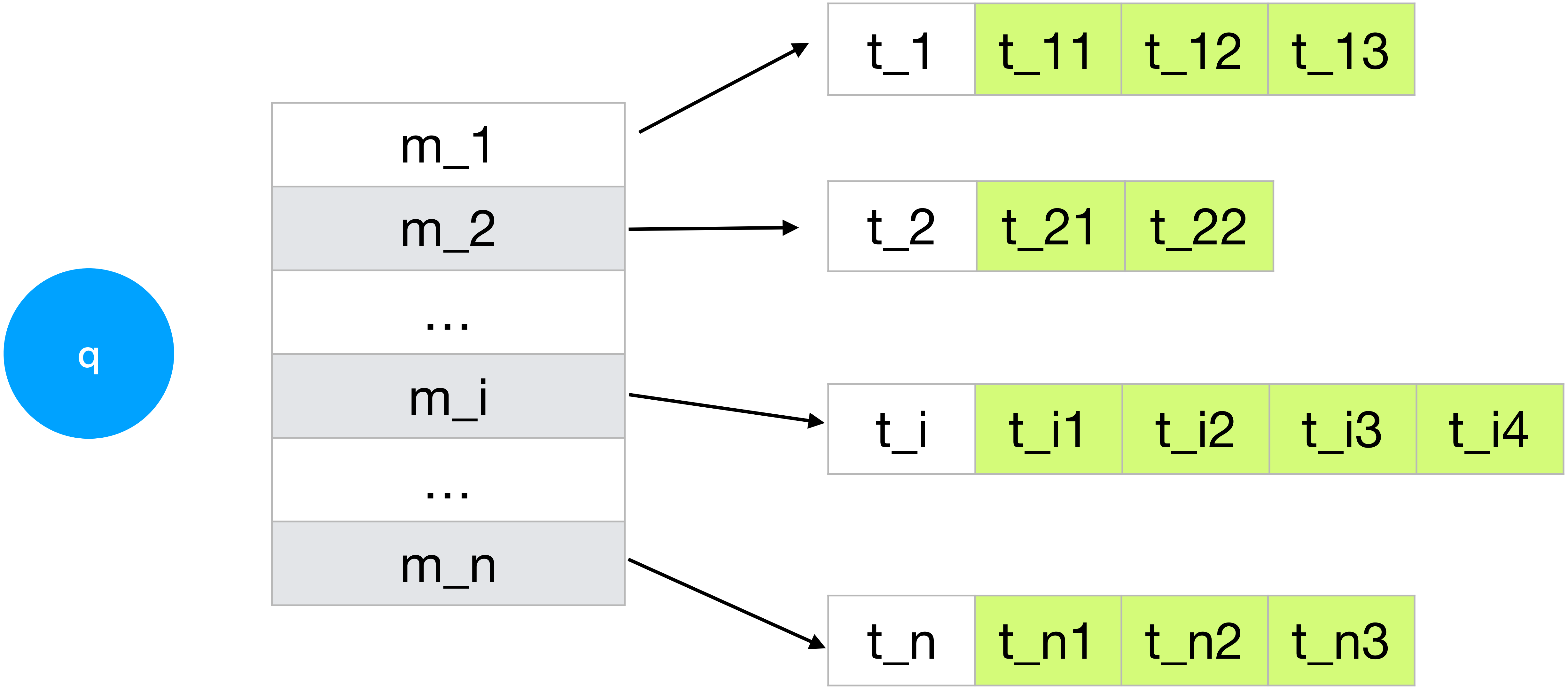
# Query expansion

- Select a subset of all the possible relations (predicates) between terms in the MeSHRDF
    - Baseline: the original query is used without any additional term.
    - Simple Expansion (SE): all the MesH entries related to that term are kept, except for the predicates `meshv:Qualifier', `meshv:seeAlso', `meshv:broader' e `meshv:broaderDescriptor' […]
    - SE + broader
    - SE + also
    - …

# Query expansion

- At the end of a query expansion process, we have
- The original query *q*
- A vector *m* = (*m_1, m_2, …, m_n*) of MeSH terms associated with the original query
- A list *t* of expanded terms (of *n* elements) where each element *t_i* is another vector of terms resulting from the iteration of the expansion approach

# Query expansion



AVIGNON CLEF 2018
Conference and Labs for the Evaluation Forum

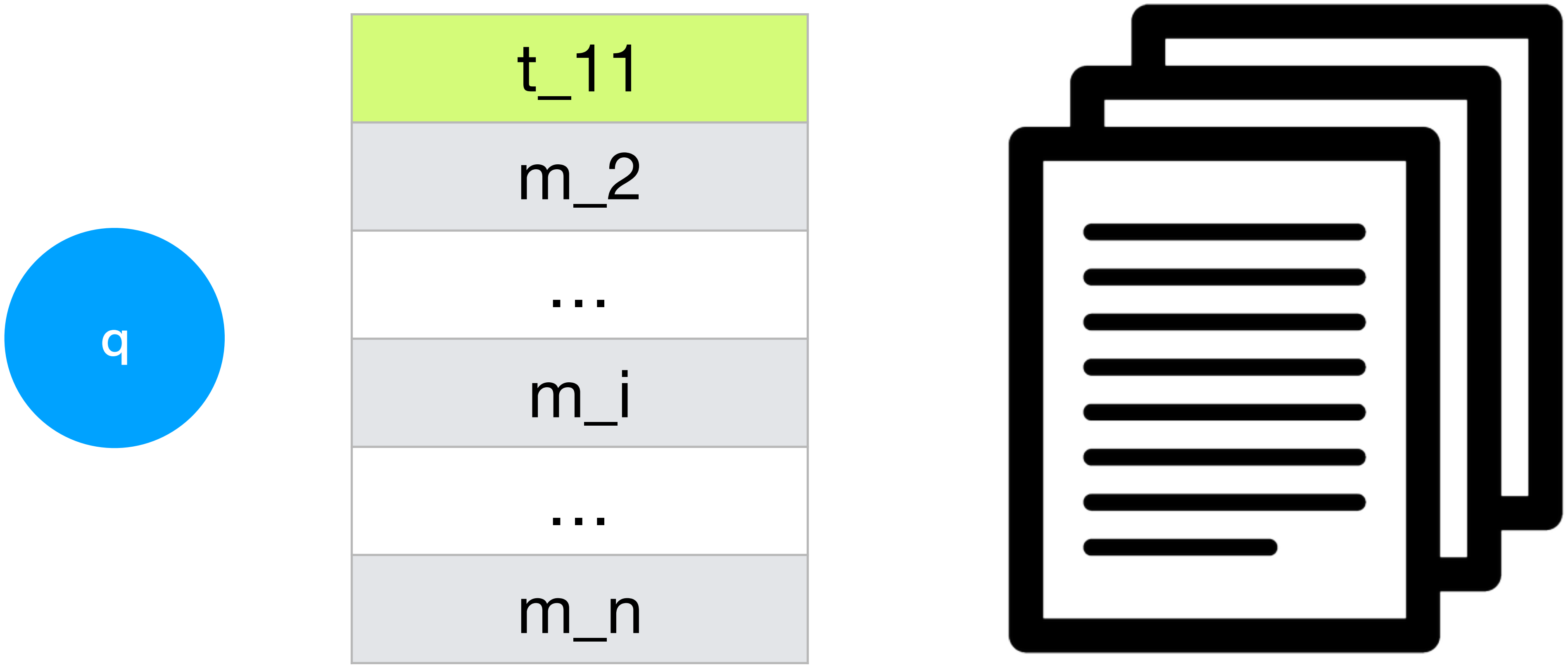
# Building query

- Give the vector of MesH terms m and the list of expanded terms t, we create a set of expanded queries by means of the following procedure:
    - For each MeSH term *m_i*
    - Substitute *m_i* with one of the terms in *t_i*, for example *t_i1*
    - Build the expanded query by merging the original query with the new set of terms *m_i** = (*m_1, …, t_i1, …, m_n*)
- At the end of the process we generate a set *V* of vectors of expanded queries where the cardinality |*V*| is the sum of all the elements in the vectors of the list t.

# Building query

| topic | simple | broader_also | recursive |
|---|---|---|---|
| average | 124.24 | 183.36 | 239.94 |

# Ranking lists



q

t_11

m_2

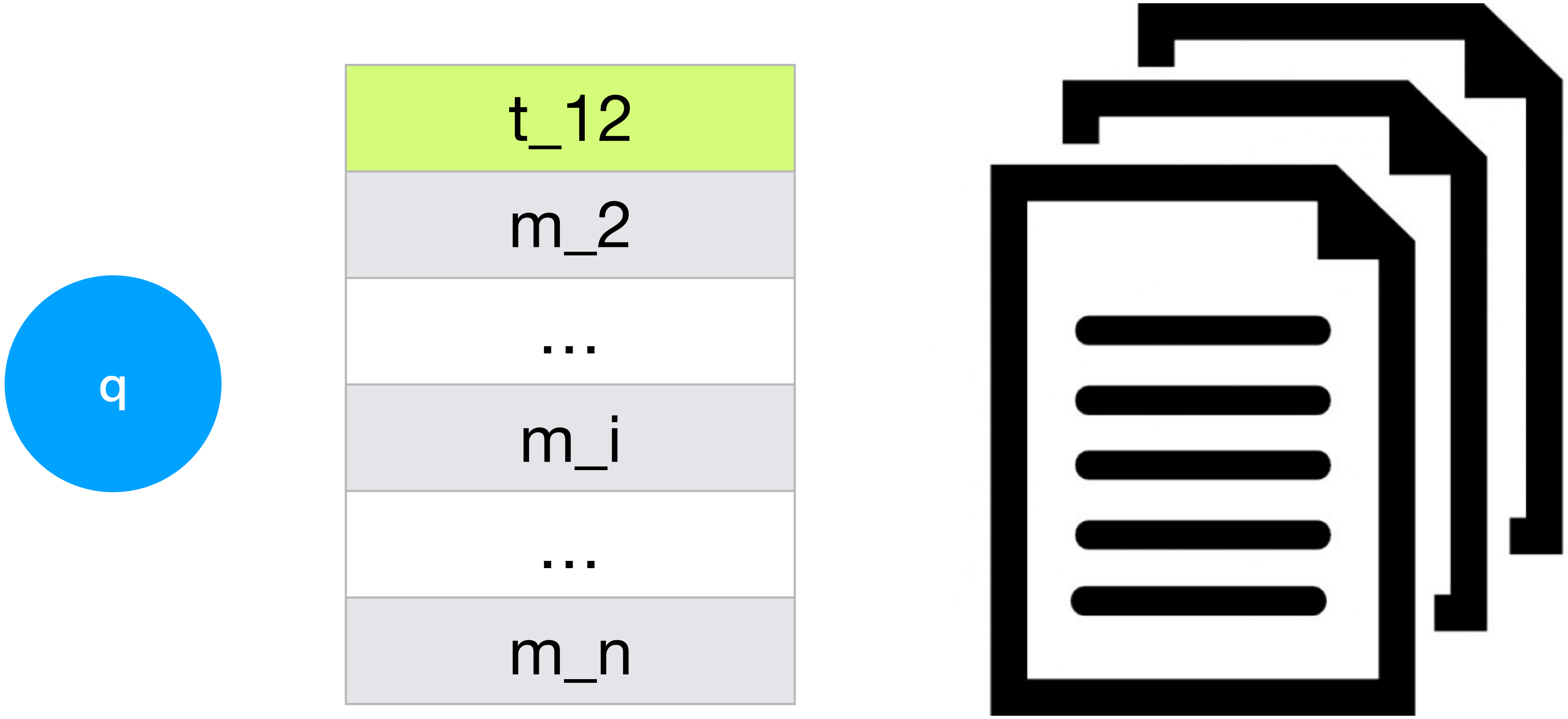
…

m_i

…

m_n

# Ranking lists



q

t_12

m_2

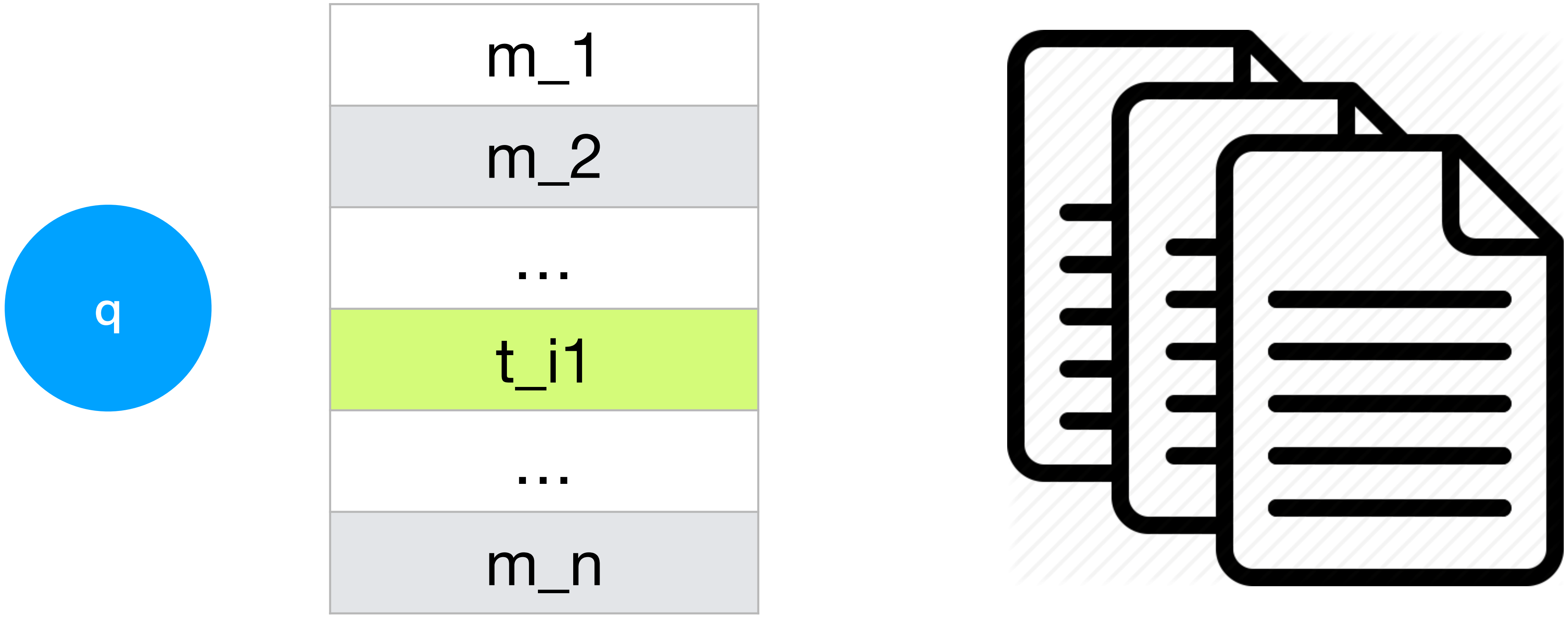
…

m_i

…

m_n

# Ranking lists

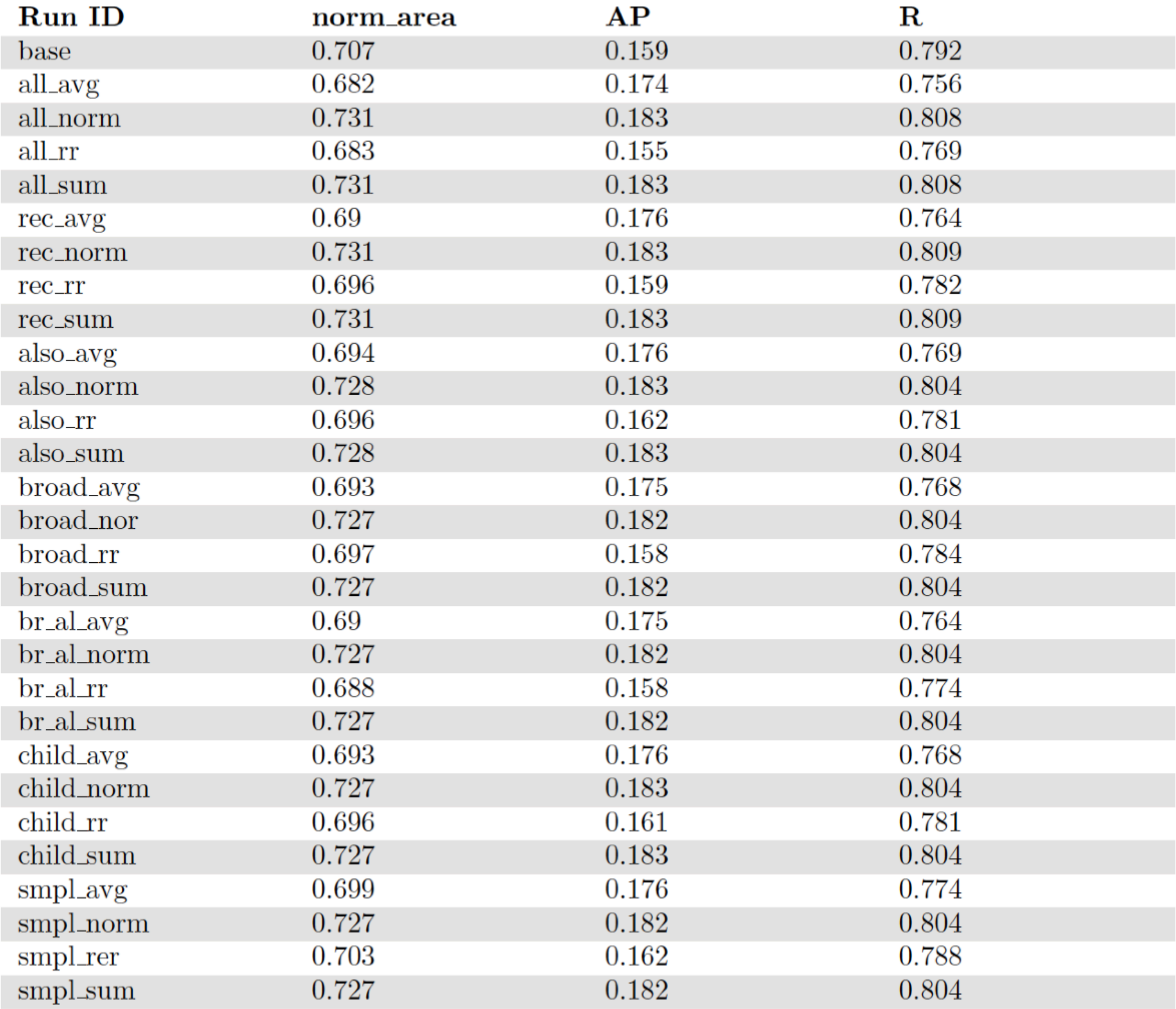


# Merging Ranking Lists

- For each expanded query we obtain a ranked list (BM25)
- Combine the scores of the (same) documents and re-rank
    - Average: given a document present in one or more lists, the scores associated to the document are averaged. Then the documents are ordered in decreasing order on the basis of this new score.
    - Sum
    - Normalized sum
    - Round robin

# Preliminary Results

- CLEF 2018 task 2 training data (42 topics)



| Run ID | norm_area | AP | R |
|---|---|---|---|
| base | 0.707 | 0.159 | 0.792 |
| all_avg | 0.682 | 0.174 | 0.756 |
| all_norm | 0.731 | 0.183 | 0.808 |
| all_rr | 0.683 | 0.155 | 0.769 |
| all_sum | 0.731 | 0.183 | 0.808 |
| rec_avg | 0.69 | 0.176 | 0.764 |
| rec_norm | 0.731 | 0.183 | 0.809 |
| rec_rr | 0.696 | 0.159 | 0.782 |
| rec_sum | 0.731 | 0.183 | 0.809 |
| also_avg | 0.694 | 0.176 | 0.769 |
| also_norm | 0.728 | 0.183 | 0.804 |
| also_rr | 0.696 | 0.162 | 0.781 |
| also_sum | 0.728 | 0.183 | 0.804 |
| broad_avg | 0.693 | 0.175 | 0.768 |
| broad_nor | 0.727 | 0.182 | 0.804 |
| broad_rr | 0.697 | 0.158 | 0.784 |
| broad_sum | 0.727 | 0.182 | 0.804 |
| br_al_avg | 0.69 | 0.175 | 0.764 |
| br_al_norm | 0.727 | 0.182 | 0.804 |
| br_al_rr | 0.688 | 0.158 | 0.774 |
| br_al_sum | 0.727 | 0.182 | 0.804 |
| child_avg | 0.693 | 0.176 | 0.768 |
| child_norm | 0.727 | 0.183 | 0.804 |
| child_rr | 0.696 | 0.161 | 0.781 |
| child_sum | 0.727 | 0.183 | 0.804 |
| smpl_avg | 0.699 | 0.176 | 0.774 |
| smpl_norm | 0.727 | 0.182 | 0.804 |
| smpl_rer | 0.703 | 0.162 | 0.788 |
| smpl_sum | 0.727 | 0.182 | 0.804 |

# Preliminary Results

- Average and Round Robin merging strategies worse than baseline (original query) for all query expansions
- Sum and Normalized Sum, the results of these merging approaches are indistinguishable
    - (probably) the large number of terms of the expanded queries make the ranking lists very similar

# Thank you!