

COVID-19 MLIA · TASK 2 MULTILINGUAL SEMANTIC SEARCH

AN INFORMATION RETRIEVAL SYSTEM FOR COVID-19 DOCUMENTS IN SPANISH

José Alberto Mesa Murgado, Pilar López Úbeda, Manuel Carlos Díaz-Galiano & María-Teresa
Martin-Valdivia
Universidad de Jaén



Universidad
de Jaén



HELLO!

ABOUT ME

My name is **José Alberto Mesa Murgado** currently working at the University of Jaén, Spain as a researcher for the group SINAI in which we are focused on the study of Human Languages Technologies (HLT)

Contact me



@murgado



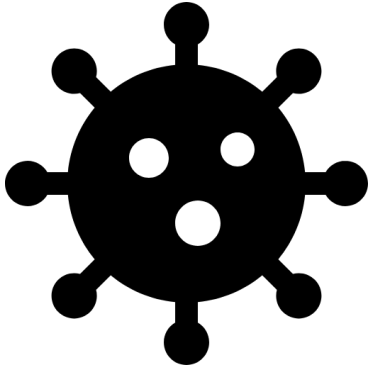
jmurgado@ujaen.es



Universidad de Jaén



Our participation



Covid-19 has arise a global concern for everyone



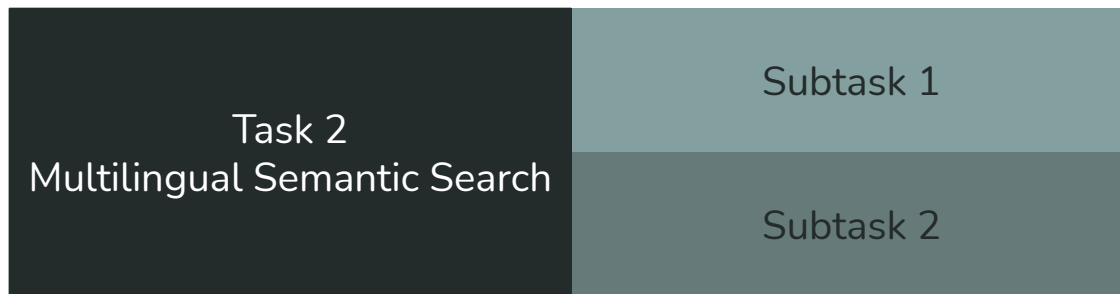
A huge number of publications have been made



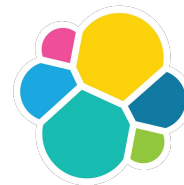
Provides an interesting challenge for us

Our participation

For our **first entry** to Covid-19 MLIA for which we inscribed in the 9th of november 2020, our initial plan was to deal with both task 1 and task 2 but after considering the time we had available we decided to only focus on the latter and both of its subtasks through a common methodology.



Methodology



elastic

Document indexation

With the Corpus in spanish, we decided to use elastic as the framework with which to index every single one of the given documents, this indexation was performed through

Document **title**

Document **keywords**

Document **content**

Run name	Components
sinai1	Search using keywords
sinai2	Search using conversational
sinai3	Search using explanation
sinai4	Search using keywords, conversational & explanation
sinai5	Search using keywords as query statement on fields: relevant keywords

Methodology

Information retrieval through elastic

Elastic implements Okapi BM25 and uses it by default for its similarity module under the parameters:

b	0.75
k1	1.2

$$\sum_i^n IDF(q_i) \frac{f(q_i, D) * (k1 + 1)}{f(q_i, D) + k1 * (1 - b + b * \frac{fieldLen}{avgFieldLen})}$$

1	Q0	medisys-es-2020_04_3337_93	0	6.3812103	sinai1
1	Q0	medisys-es-2020_03_19723_20	1	6.379256	sinai1
1	Q0	medisys-es-2020_03_2265_191	2	6.36521	sinai1
...

```
File Edit Selection View Go Run Terminal Help findDocs.py - TASK 2 - Visual Studio Code
findDocs.py x
Scripts > findDocs.py
# Retrieve documents from elasticsearch based on the concept more like this
# https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mit-query.html
# query: Textual query
def getDocuments(self, topic_query, searchByTitle = False, searchByKeyword = False, searchByContent = False):
    response = False

    # Query format
    body = { 'query': {} }

    if topic_query:
        # Elasticsearch Query Format transformation

        body['query'] = {
            'more_like_this': {
                'fields': [],
                'like': topic_query,
                'min_term_freq': 1,
                'max_query_terms': 20,
            }
        }

    # Restrict query
    if not searchByTitle and not searchByKeyword and not searchByContent:
        raise Exception('Not field provided to perform document retrieval')

    if searchByTitle:
        body['query']['more_like_this']['fields'].append('title')

    if searchByKeyword:
        body['query']['more_like_this']['fields'].append('keyword')

    if searchByContent:
        body['query']['more_like_this']['fields'].append('content')

    # Call Elasticsearch and format given response
    try:
        response = []
        returned = self.es.search(body=body, index=INDEX, size=size)
        for h in returned['hits']['hits']:
            item = { 'score': h['_score'], 'id': h['_id'], 'title': h['_source']['title'] }
            response.append(item)
    except Exception:
        pass

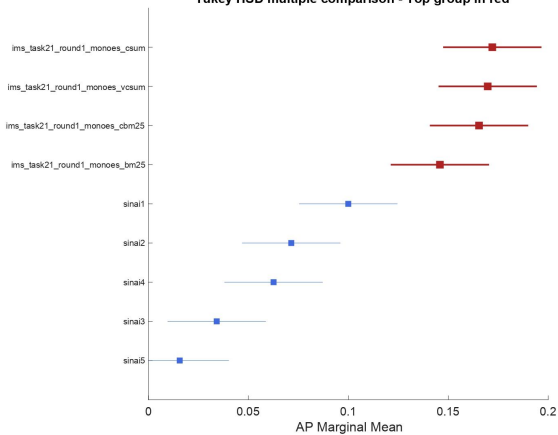
    return response
```

Results I

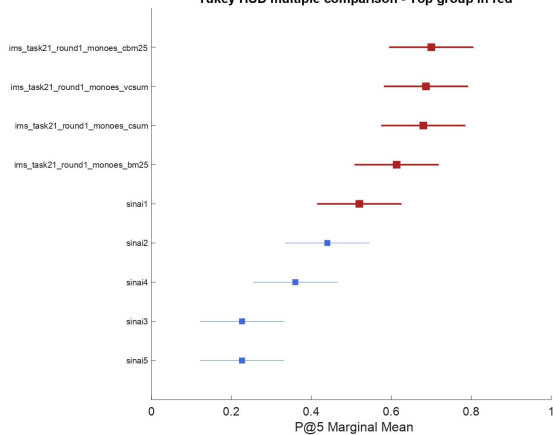
Task 2 · Subtask 1

High Precision

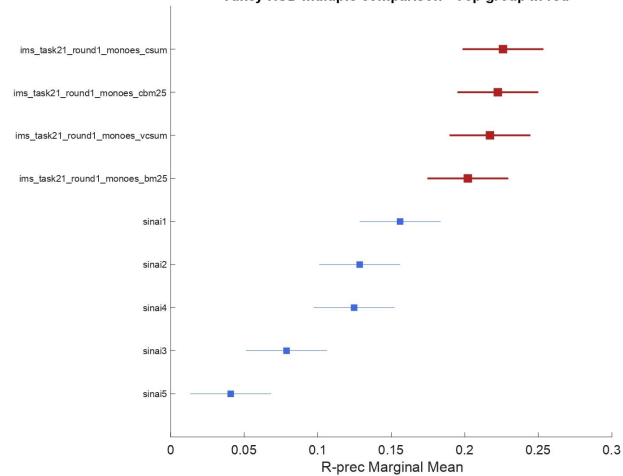
Tukey HSD multiple comparison - Top group in red



Tukey HSD multiple comparison - Top group in red



Tukey HSD multiple comparison - Top group in red

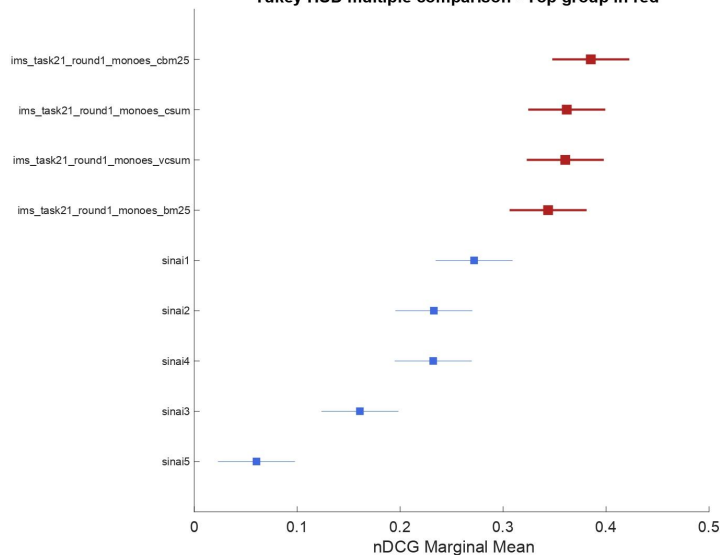


Results I

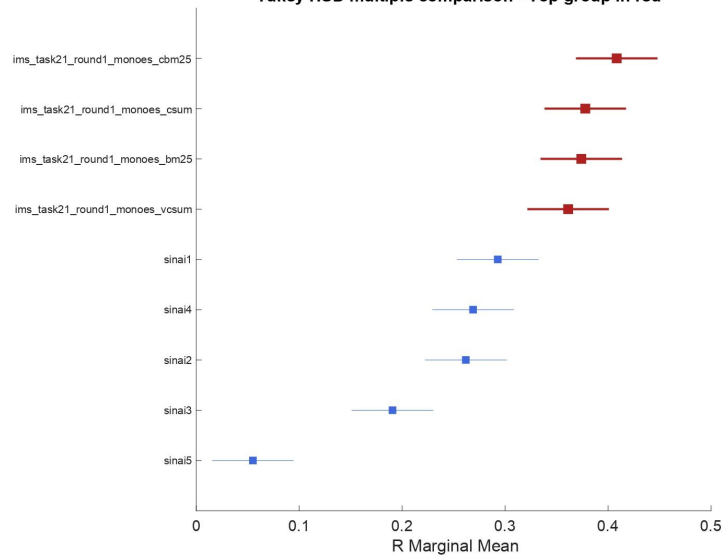
Task 2 · Subtask 1

High Precision

Tukey HSD multiple comparison - Top group in red



Tukey HSD multiple comparison - Top group in red

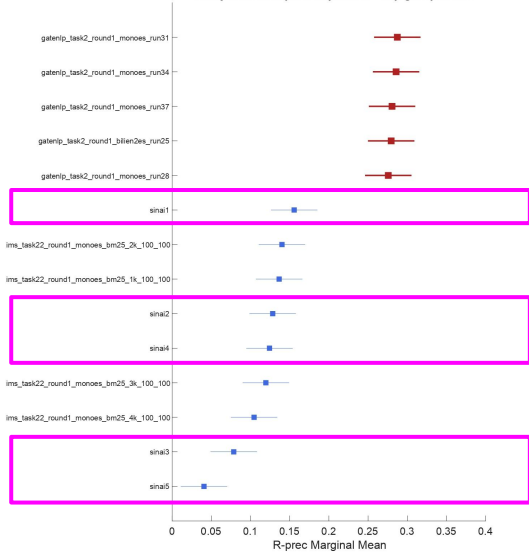


Results II

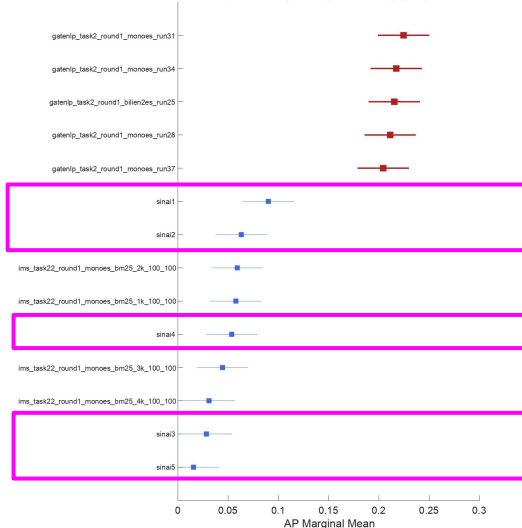
Task 2 · Subtask 2

High Recall

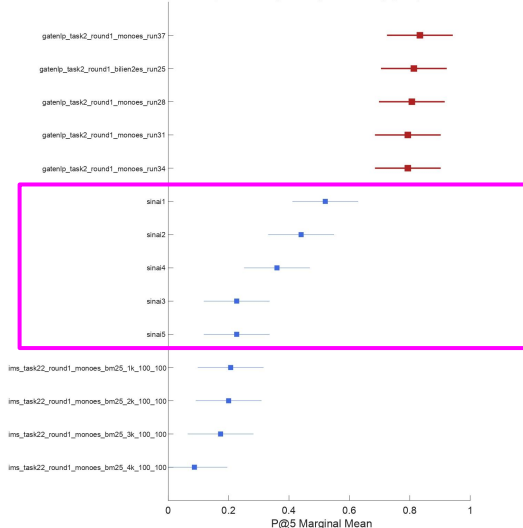
Tukey HSD multiple comparison - Top group in red



Tukey HSD multiple comparison - Top group in red



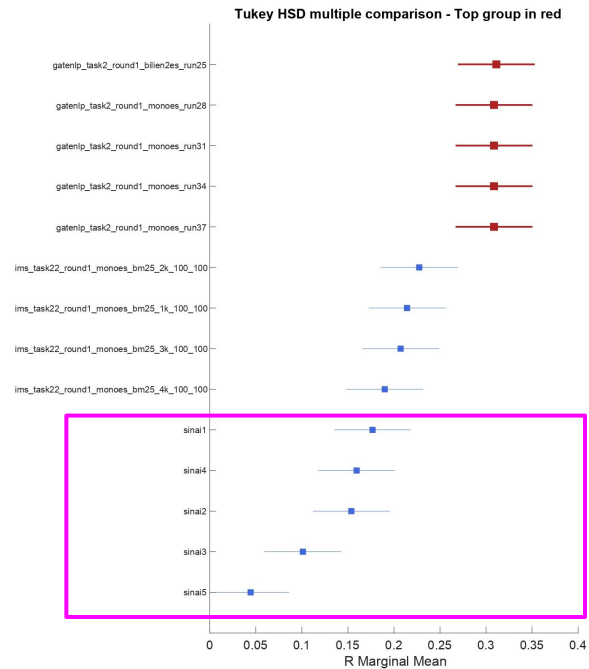
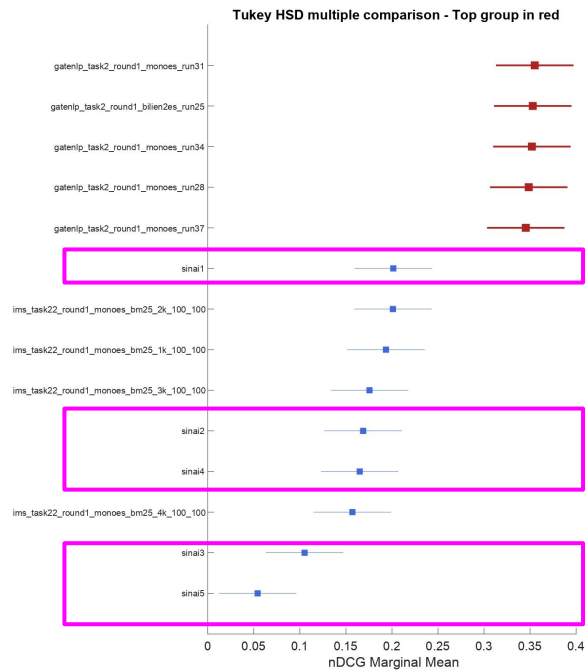
Tukey HSD multiple comparison - Top group in red



Results II

Task 2 · Subtask 2

High Recall



Discussion & Future work

Discussion

This is a **preliminary approach** for our participation on the initiative which we will improve

Shortage of participants means that no major comparisons can be made

We have to make **improvements on our systems in terms of precision and recall** with an emphasis on the latter.

Future approach

Identify entities through NER



Improve our queries

AN INFORMATION RETRIEVAL SYSTEM FOR COVID-19 DOCUMENTS IN SPANISH

José Alberto Mesa Murgado · Universidad de Jaén

Contact me



@murgado



jmurgado@ujaen.es



Universidad
de Jaén

