Συστήματα Ανάκτησης Πληροφοριών

Προγραμματιστική Εργασία: Μέρος Β

Σοφία-Ζωή Σωτηρίου, p3210192

Πριν το περιεχόμενο του δεύτερου ερωτήματος, αναφέρεται ότι εκπρόθεσμα σε σχέση με τη διορία του πρώτου μέρους (και άρα γι' αυτό αναφέρεται σε αυτό το σημείο) δοκιμάστηκε ένας νέος τρόπος απλής αναζήτησης χωρίς συνώνυμα, ο οποίος απέφερε καλύτερα αποτελέσματα από αυτά που καταγράφονται στο πρώτο μέρος, οπότε και στο δεύτερο μέρος εφαρμόστηκε αυτός.

Πιο συγκεκριμένα, δοκιμάστηκε η αναζήτηση με τη χρήση του πεδίου allContent σε σχέση με όλα τα πεδία ενός query. Ο κώδικας που χρησιμοποιήθηκε είναι ο εξής:

```
with open("write/results.test", "w") as file :
    for query in formatted_queries :
        search_results = client.search(
            index="indexino",
            size=20,
            query={
                "match": {"allContent": query.get("text", "")},
                "match": {"allContent": safe_query(query.get("authors", ""))},
                "match": {"allContent": safe_query(query.get("authors", ""))},
                "match": {"allContent": query.get("year", "")},
                "match": {"allContent": safe_query(query.get("cited_by", ""))},
                "match": {"allContent": safe_query(query.get("references", ""))},

            }
        )
        c = 1
        for hit in search_results["hits"]["hits"] :
            doc_id = hit["_id"]
            score = hit.get("score", None)
            run_name = "STANDARD"
            rank = c
            c+=1

            line = f"{query["uniqueID"]} Q0 {doc_id} {rank} {score} {run_name}"
            file.write(line + "\n")
✓ 21.6s
```

Ο οποίος επέφερε τα εξής αποτελέσματα (για size=20) στο ευρετήριο indexino που χρησιμοποιήθηκε και στο προηγούμενο σκέλος της εργασίας:

```
num_rel                 all     4928
num_rel_ret             all     2309
map                     all     0.1616
gm_map                  all     0.0176
```

```
P_5                     all     0.1362
P_10                    all     0.1330
P_15                    all     0.1213
P_20                    all     0.1155
```

Τα οποία είναι εμφανώς καλύτερα από τα προηγούμενα καλύτερα αποτελέσματα:

```
num_rel                 all     4908
num_rel_ret             all     1692
map                     all     0.0955
gm_map                  all     0.0091
```

```
P_5                     all     0.0900
P_10                    all     0.0889
P_15                    all     0.0859
P_20                    all     0.0849
```

Ε

Επομένως, για την συνέχεια της εργασίας κρατείται αυτός ο κώδικας που αποφέρει:

Για size=20

```
num_rel                 all     4928
num_rel_ret             all     2309
map                     all     0.1616
gm_map                  all     0.0176
```

```
P_5                     all     0.1362
P_10                    all     0.1330
P_15                    all     0.1213
P_20                    all     0.1155
```

Για size=30

```
num_rel                 all     4928
num_rel_ret             all     2586
map                     all     0.1634
gm_map                  all     0.0258
```

```
P_5                     all     0.1340
P_10                    all     0.1279
P_15                    all     0.1138
P_20                    all     0.1010
```

Για size=50

```
num_rel                 all     4928
num_rel_ret             all     2851
map                     all     0.1647
gm_map                  all     0.0356
```

```
P_5                     all     0.1344
P_10                    all     0.1265
P_15                    all     0.1115
P_20                    all     0.0970
```

Για το δεύτερο μέρος της εργασίας, προτιμήθηκε ο πρώτος τρόπος, δηλαδή η χρήση synonym_graph_filter στον αναλυτή της elasticSearch.

Επομένως, άλλαξε ο ήδη υπάρχων κώδικας, τροποποιώντας το mapping ώστε να προστεθεί το επιθυμητό φίλτρο και analyzer, το οποίο επιτεύχθηκε με τη παρακάτω προσθήκη:

```
"analysis": {
    "filter": {
        "wn_synonym_filter": {
            "type": "synonym_graph",
            "format" : "wordnet",
            "synonyms_path": "read/wn_s_nouns_verbs.pl"
        },
        "english_stop": {
            "type": "stop",
            "stopwords": "_english_"
        },
        "english_stemmer": {
            "type": "stemmer",
            "language": "english"
        }
    },
    "analyzer": {
        "default": {
            "tokenizer": "standard",
            "filter": [
                "lowercase",
                "english_stemmer",
                "wn_synonym_filter",
                "english_stop"
            ]
        }
    }
}
```

Τα σενάρια που επιλέχθηκαν ήταν η επέκταση με συνώνυμα των ουσιαστικών (αρχείο wn_s_nouns.pl) και η επέκταση με συνώνυμα των ουσιαστικών και των ρημάτων (αρχείο wn_s_nouns_verbs.pl). Για τον σκοπό αυτό δημιουργήθηκαν 2 νέα ευρετήρια, το nouns και το nv αντίστοιχα.

Τα αποτελέσματα με την επέκταση με συνώνυμα μόνο των ουσιαστικών έχουν ως εξής:

Για k=20 (όπως φαίνεται και στο αρχείο nouns20)

num_rel	all	4928
num_rel_ret	all	2534
map	all	0.1700
gm_map	all	0.0345

P_5	all	0.1422
P_10	all	0.1406
P_15	all	0.1309
P_20	all	0.1267

Για k=30 (όπως φαίνεται στο αρχείο nouns30)

num_rel	all	4928
num_rel_ret	all	2771
map	all	0.1689
gm_map	all	0.0441

P_5	all	0.1364
P_10	all	0.1316
P_15	all	0.1181
P_20	all	0.1060

Για k=50 (όπως φαίνεται στο αρχείο nouns50)

num_rel	all	4928
num_rel_ret	all	3001
map	all	0.1677
gm_map	all	0.0504

P_5	all	0.1360
P_10	all	0.1276
P_15	all	0.1129
P_20	all	0.0985

Αποτελέσματα που προέκυψαν με την επέκταση με συνώνυμα των ουσιαστικών και των ρημάτων:

Για k=20 (όπως φαίνεται και στο αρχείο nv20)

num_rel	all	4928
num_rel_ret	all	2569
map	all	0.1721
gm_map	all	0.0377

P_5	all	0.1436
P_10	all	0.1423
P_15	all	0.1323
P_20	all	0.1285

Για k=30 (όπως φαίνεται και στο αρχείο nv30)

num_rel	all	4928
num_rel_ret	all	2801
map	all	0.1699
gm_map	all	0.0464

P_5	all	0.1376
P_10	all	0.1319
P_15	all	0.1189
P_20	all	0.1063

Για k=50 (όπως φαίνεται και στο αρχείο nv50)

num_rel	all	4928
num_rel_ret	all	3027
map	all	0.1681
gm_map	all	0.0527

P_5	all	0.1360
P_10	all	0.1278
P_15	all	0.1134
P_20	all	0.0988

Αποτελέσματα ανά μετρική

*Για την αναφορά στα τρία σενάρια, θα χρησιμοποιούνται τα ευρετήρια που χρησιμοποιήθηκαν σε κάθε περίπτωση, δηλαδή indexino-χωρίς επέκταση, nouns-επέκταση με συνώνυμα των ουσιαστικών, nv-επέκταση με συνώνυμα και ουσιαστικών και ρημάτων.

Αριθμός σχετικών ανακτηθέντων κειμένων από τα 4928 που έπρεπε (num_rel_ret)

	k=20	k=30	k=50
indexino	2309	2586	2851
nouns	2534	2771	3001
nv	2569	2801	3027

Mean Average Precision (MAP)

	k=20	k=30	k=50
indexino	0.1616	0.1634	0.1647
nouns	0.1700	0.1689	0.1677
nv	0.1721	0.1699	0.1681

Geometric Mean Average Precision (gmMAP)

	k=20	k=30	k=50
indexino	0.0176	0.0258	0.0356
nouns	0.0345	0.0441	0.0504
nv	0.0377	0.0464	0.0527

Precision στα 5 πρώτα ανακτηθέντα κείμενα (P5)

	k=20	k=30	k=50
indexino	0.1362	0.1340	0.1344
nouns	0.1422	0.1364	0.1360
nv	0.1436	0.1376	0.1360

Precision στα 10 πρώτα ανακτηθέντα κείμενα (P10)

	k=20	k=30	k=50
indexino	0.1330	0.1279	0.1265
nouns	0.1406	0.1316	0.1276
nv	0.1423	0.1319	0.1278

Precision στα 15 πρώτα ανακτηθέντα κείμενα (P15)

	k=20	k=30	k=50
indexino	0.1213	0.1138	0.1115
nouns	0.1309	0.1181	0.1129
nv	0.1323	0.1189	0.1134

Precision στα 20 πρώτα ανακτηθέντα κείμενα (P20)

	k=20	k=30	k=50
indexino	0.1155	0.1010	0.0970
nouns	0.1267	0.1060	0.0985
nv	0.1285	0.1063	0.0988

Συμπεράσματα

Αρχικά, πρέπει να σημειωθεί ότι τα δυο σενάρια (nouns και nv) επιλέχθηκαν με τη λογική ότι σε κάθε σενάριο θα υπάρχει κάποια κλιμάκωση (από κανένα συνώνυμο στα ουσιαστικά και από εκεί στα ουσιαστικά και ρήματα) ώστε να φανεί αν με τη κάθε επέκταση υπήρξε και βελτίωση των αποτελεσμάτων αντί να γίνει απλά σύγκριση απόδοσης δύο ξένων μεταξύ τους σεναρίων, όπως θα γινόταν αν π.χ. είχαν επιλεχθεί τα συνώνυμα των ουσιαστικών και των ρημάτων.

Σχετικά με την ανάκτηση σχετικών κειμένων, είναι φανερό ότι ανεξάρτητα από το k (αριθμός των ανακτηθέντων κειμένων - 20, 30 ή 50), με το nouns αυξήθηκαν σημαντικά τα σχετικά κείμενα (~180 παραπάνω σχετικά κείμενα) ενώ με το nv τα αυξήθηκαν, αλλά σε πολύ μικρότερο βαθμό (~25 παραπάνω σχετικά κείμενα). Έτσι φάνηκε ότι η αρχική επέκταση βοηθάει στην ουσιαστική βελτίωση της ανάκτησης, ενώ οι περαιτέρω προσθήκες αποφέρουν μικρότερες βελτιώσεις. Το μοτίβο αυτό εμφανίζεται σε όλες τις μετρικές, με το nouns να φέρνει μια σημαντική βελτίωση και το nv να φέρνει βελτίωση, αλλά εμφανώς μικρότερη. Αλλά σε κάθε περίπτωση, τα καλύτερα αποτελέσματα τα φέρνει πάντα το nv, δηλαδή η μεγαλύτερη συνολικά επέκταση.

Επίσης, όπως είναι λογικό, όσο αυξάνεται το k αυξάνεται και ο αριθμός των σχετικών ανακτηθέντων κειμένων (και στα τρία σενάρια, για k=50 βρέθηκαν συνολικά ~500 παραπάνω σχετικά κείμενα σε σχέση με το k=20). Ωστόσο, όσο αυξάνεται το k, η αύξηση των σχετικών ανακτηθέντων κειμένων μικραίνει, δηλαδή για k=20 οι αυξήσεις από σενάριο σε σενάριο ήταν 223 και 37 έξτρα κείμενα, ενώ για k=50 ανακτήθηκαν 153 και έπειτα 23 έξτρα κείμενα, επιβεβαιώνοντας ότι οι βελτιώσεις στην αναζήτηση επηρεάζουν περισσότερο τα πρώτα ανακτηθέντα κείμενα. Αυτό το μοτίβο επαναλαμβάνεται σε όλες τις μετρικές, δηλαδή όσο μεγαλώνει το k τόσο μικραίνουν οι βελτιώσεις.

Το MAP έχει τη καλύτερη τιμή του για k=20 και ευρετήριο το nv, δείχνοντας ότι τα περισσότερα σχετικά κείμενα βρίσκονται ψηλά στη λίστα των ανακτηθέντων (άρα όταν ανακτώνται περισσότερα κείμενα είναι λιγότερο σχετικά). Επίσης, υπάρχουν και τα προαναφερθείσα μοτίβα, δηλαδή ότι όσο αυξάνεται το k μειώνονται οι αλλαγές, και ότι από σενάριο σε σενάριο η συνολική βελτίωση μικραίνει.

Σχετικά με το gmMAP, η σημαντική του βελτίωση από το indexino στο nouns για όλες τις τιμές του k δείχνει το πόσο σταθεροποιεί την αναζήτηση η επέκταση με συνώνυμα και πόσο βοηθάει στο να απαντηθούν «δύσκολα» ερωτήματα, ειδικά στα πρώτα κείμενα, καθώς για k=20 το gmMAP πρακτικά διπλασιάζεται. Αυτό σημαίνει ότι πλέον και στα πιο «δύσκολα» ερωτήματα βρίσκονται σχετικές απαντήσεις με τη χρήση συνωνύμων, με τα καλύτερα αποτελέσματα αυτή τη φορά να είναι στο nv, με k=50 όμως, που δείχνει ότι βελτιώνεται η συνολική απόδοση το συστήματος, καθώς και για πολλά ανακτηθέντα κείμενα υπάρχουν σχετικές απαντήσεις.

Το precision των πρώτων ανακτηθέντων κειμένων είναι καλύτερο στα πέντε πρώτα ανακτηθέντα κείμενα στην αναζήτηση στο nv, δείχνοντας ότι η καλύτερη αναζήτηση γίνεται στην επέκταση και με ουσιαστικά και με ρήματα, αλλά και ότι τα πιο σχετικά κείμενα είναι τα πρώτα, ενώ μετά στα επόμενα ανακτηθέντα κείμενα (10 πρώτα, 15 πρώτα και 20 πρώτα) το αντίστοιχο precision πέφτει σταθερά για κάθε συνδυασμό σεναρίου και k. Επίσης, φαίνεται ότι το precision πέφτει και όσο αυξάνονται να ανακτηθέντα κείμενα, καθώς σε όλους τους πίνακες και για όλα τα σενάρια το precision για k=20 είναι το μεγαλύτερο.

Τέλος, σημειώνεται ότι από τη φάση 1 στη φάση 2 της εργασίας τα αποτελέσματα έχουν βελτιωθεί σημαντικά, το οποίο είναι λογικό καθώς στη φάση 2 πέρα από word-for-word αναζήτηση, πραγματοποιείται και εννοιολογική αναζήτηση μέσω των συνωνύμων.