

Στο τρίτο και τελευταίο μέρος της εργασίας χάρης στις τεχνικές που χρησιμοποίησα παρακάτω κατάφερα να βελτιώσω σημαντικά τα αποτελέσματα από το 2^ο και το 1^ο κομμάτι της εργασία βρίσκοντας συνολικά **99** relative retrieved στα πρώτα 50 του κάθε query.

Για την εκτέλεση της εργασίας χρησιμοποίησα τον κώδικα του 6^{ου} εργαστηρίου έκανα αρχικά import στο intelliJ όλο το project, όμως επειδή δεν με βοηθούσε ο κώδικας και με υπήρχαν προβλήματα με τις εκδόσεις, επιπλοκές μεταξύ τους και δεν αναγνώριζε τους analysers της Lucene και τις κλάσεις του πχ. Stop_Words αποφάσισα μετά από πολύ αναζήτηση στο διαδίκτυο να το χτίσω όλο από την αρχή με τον κώδικα δημιουργίας ευρετηρίου της 1^{ης} εργασίας και τον αντίστοιχο Searcher όπως ανέφερε η καθηγήτρια στην περιοχή συζητήσεων'.

Για το word2vec χρησιμοποίησα τον κώδικα του 6^{ου} εργαστηρίου ο οποίος με κάποιες αλλαγές δημιουργεί πλέον ένα αρχείο w2vec.txt που εκεί αποθηκεύει την λέξη και κατόπιν τα συνώνυμα που βρέθηκαν.

Τα αρχικά αποτελέσματα ήταν αρκετά απογοητευτικά καθώς έλαβα μόλις 57 relative retrieved κείμενα και με διάφορους πειραματισμούς και διάβασμα και από το διαδίκτυο και από την περιοχή συζητήσεων κατέληξα σε μία πολύ χρονοβόρα εκπαίδευση 2.5 ωρών με τις παρακάτω ρυθμίσεις στον w2vec builder.

```
org.deeplearning4j.models.word2vec.Word2Vec vec = new
org.deeplearning4j.models.word2vec.Word2Vec.Builder()
    .minWordFrequency(4)
    .layerSize(400)//TODO
    .seed(42)
    .windowSize(20)//TODO
    .epochs(10)
    .elementsLearningAlgorithm(MLP)
    .tokenizerFactory(new LuceneTokenizerFactory(new WhitespaceAnalyzer()))//TODO
    .iterate(iter)
    .build();
```

- Άλλαξα το minWordFrequency σε 4 από 5 ώστε να μειωθούν οι απαιτούμενες εμφανίσεις ενός όρου προκειμένου να γίνει vector,
- Άλλαξα layerSize(400) από το αρχικό και default 100 με αποτέλεσμα η πολυπλοκότητα και το μέγεθος του μοντέλου εκπαίδευσης να αυξηθεί αρκετά και το κράτησα πολλαπλάσιο του 4 καθώς στο διαδίκτυο αναφέρουν πως βοηθάει.
- Επίσης «μείωσα» την αρκίβεια του μοντέλου και λαμβάνει πλέον υπόψη του όχι τις προεπιλεγμένες 5 λέξεις «υπόψη» του αλλά τις 20 προκειμένου να βγουν πιο ολοκληρωμένα vectors.
- Άλλαξα τον analyser από τον default του εργαστηρίου Standard σε EnglishAnalyzer αλλά επειδή τα αποτελέσματα ήταν τραγικά (έβρισκε περίπου 5 συνώνυμα για όλους τους όρους συνολικά) δοκίμασα τον WhitespaceAnalyzer ο οποίος έβρισκε για όλους τους όρους αποτελέσματα κάτι το οποίο δεν πετύχαινε ούτε ο Standard καθώς δεν έβρισκε σχεδόν ποτέ για τους όρους cross-domain και Mobility-as-a-Service.

• Το αρχείο εκπαίδευσης ήταν αυτό που ανέβασε ο κ. Μοσχοβής με όνομα IR2020_clean.txt στην περιοχή χωρίς καμία τροποποίηση.

• Οι όροι που πραγματοποίησα την αναζήτηση ήταν σχεδόν όλοι οι όροι των queries γιατί θα τους τροποποιούσα αργότερα εκτός φυσικά αυτών που κοβόντουσαν από τον EnglishAnalyzer px (of for etc)

Οι όροι λοιπόν για τους οποίους εκπαιδεύτηκε αρχικά το μοντέλο προκειμένου να παράξει παράγωγα ήταν οι παρακάτω σε lowercase για να εντοπιστούν στο αρχείο εκπαίδευσης :

"multimodal", "travel", "services", "big", "data", "mobility", "european", "logistics", "applications", "architectures", "big", "data", "analytics", "architecture", "industrial", "iot", "mobility-as-a-service", "tools", "fragmentation", "iot", "federation", "seamless", "efficient", "european", "travelling", "cross-domain", "orchestration", "services", "community", "networks"

• Για κάθε όρο αποθήκευσα στο αρχείο αρχικά 15 συνώνυμα τα οποία ήταν τα εξής:

```
whitespace - L:400 window:20 min:4 words:15
multimodal [multi-modal, modalities, neuroimaging, imaging, perception, speech, fbi, cognitive, stimulation, tactile, facial, modality, virtual, sensorimotor, fmri]
travel [passengers, trip, ticketing, door-to-door, destination, booking, travellers, passenger, dora, journey, traffic, seamless, traveller, trips, mobility]
services [service, support, management, smes, providers, innovation, users, provision, business, access, infrastructure, manage, enterprise, smart, ecosystem]
big [analytics, data, data-driven, cloud, intelligence, computing, collection, privacy, mining, bigmedilitics, cloud-based, queries, veracity, unstructured, companies]
data [information, big, analytics, analysis, datasets, provide, tools, collection, collected, methods, models, use, using, algorithms, statistical]
mobility [urban, cities, city, congestion, commuters, transport, transportation, people, travel, personal, cars, parking, destination, car, vehicle]
european [europe, eu, international, countries, national, innovation, industry, union, initiatives, support, organisations, commission, europes, policy, global]
logistics [logistic, transportation, freight, goods, intermodal, last-mile, chain, shipping, planning, supply, providers, cargo, chains, automation, operations]
applications [application, devices, technologies, materials, systems, advanced, technology, electronics, high, design, functionalities, industrial, technological, range, such]
architectures [reconfigurable, computing, architecture, heterogeneous, scalable, scalability, exascale, low-power, hardware, cross-layer, multi-core, programmability, programmable, embedded, hybrid]
big [analytics, data, data-driven, cloud, intelligence, computing, collection, privacy, mining, bigmedilitics, cloud-based, queries, veracity, unstructured, companies]
data [information, big, analytics, analysis, datasets, provide, tools, collection, collected, methods, models, use, using, algorithms, statistical]
analytics [big, data, real-time, intelligence, algorithms, cloud, cloud-based, software, online, unstructured, intelligent, hadoop, leveraging, data-driven, queries]
architecture [architectures, hardware, programmability, scalable, components, embedded, programmable, integration, reconfigurable, orchestration, architectural, virtualization, modular, computing, heterogeneous]
industrial [industry, industries, partners, sectors, manufacturing, consortium, products, sector, production, academic, companies, automotive, commercial, academia, technologies]
iot [things, internet, smart, cloud, platforms, connected, interoperability, services, ict, end-to-end, security, secure, mobile, interoperable, developers]
mobility-as-a-service [my-trac, car-sharing, last-mile, crowd-sourced, seamless, maas, av-ready, mycorridor, sharing, services, arcadia, city, socialcar, keyless, service]
tools [methods, techniques, approaches, develop, models, tool, provide, analysis, methodologies, framework, developing, modelling, existing, providing, advanced]
fragmentation [fragmented, ecosystem, synergies, bilbao, fragclm, coalescence, mass, atacama, ensuf, era-min, observatory, biodiversity, observations, initiatives, supernova]
iot [things, internet, smart, cloud, platforms, connected, interoperability, services, ict, end-to-end, security, secure, mobile, interoperable, developers]
federation [federated, fedfire, cloud, testbeds, interoperability, roaming, end-to-end, iot, interoperation, open, orchestration, federations, infrastructure, testbed, deployment]
seamless [interoperability, interoperable, user, end-to-end, users, services, sharing, service, providers, secure, platform, mobile, trusted, user-centric, privacy]
efficient [efficiency, solutions, technologies, generation, highly, sustainable, effective, developing, cost-effective, design, cost, integrated, greener, energy, systems]
european [europe, eu, international, countries, national, innovation, industry, union, initiatives, support, organisations, commission, europes, policy, global]
travelling [travellers, travel, ipco, door-to-door, openings, mobility, cloudfone, nuisance, enclosures, trip, seamless, ionosphere, home, fires, travelers]
cross-domain [interoperability, orchestration, agile, instantiate, 5gex, multi-vendor, iot, sonata, nfv, end-to-end, multi-domain, virtualized, interoperable, middleware, multi-tenancy]
orchestration [end-to-end, virtualization, nfv, virtualised, multi-tenancy, 5g-ready, 5gpagoda, virtualized, multi-operator, programmability, 5g, instantiate, cloud, multi-tenant, federated]
services [service, support, management, smes, providers, innovation, users, provision, business, access, infrastructure, manage, enterprise, smart, ecosystem]
community [communities, stakeholders, engagement, initiatives, activities, engage, makers, ecosystem, policy, public, outreach, foster, dialogue, open]
networks [network, communication, communications, wireless, 5g, connectivity, distributed, nodes, in-network, internet, virtualization, caching, sdn, networking, c-ran]
```

• Τέλος πρόσθεσα βασισμένος στον κώδικα του εργαστηρίου του αρχείου pom τα dependencies που θα χρησιμοποιούσα και χειρωνακτικά στα libraries τα αρχεία της lucene που θα χρησιμοποιούσα.

Στον Indexer και στον Searcher που χρησιμοποιήθηκαν για την δημιουργία ευρετηρίου και την πραγματοποίηση queries σε αυτόν χρησιμοποιήθηκε ο English Analyzer καθώς παρήγαγε τα καλύτερα αποτελέσματα.

Ο Indexer ο Trainer που εκπαιδεύει το μοντέλο w2vec και παράγει τα συνώνυμα καθώς και ο Searcher τρέχουν από μία κεντρική main.

```
public class Main{
    public static void main(String[] args) {
        try {
            Indexer indexerDemo = new Indexer();
            Trainer trainer = new Trainer(); // TODO
            Searcher searcher = new Searcher();
            System.out.println("\n\tResults are ready");
        } catch (Exception ex) {
            ex.printStackTrace();
        }
    }
}
```

Ο Searcher ο οποίος ήταν και αυτός ο οποίος έτρεχα περισσότερο για την βελτίωση και τον έλεγχο των αποτελεσμάτων αφού εκπαίδευσα το μοντέλο και την έβγαλα τη λίστα των συνωνύμων διαβάζει το αρχείο w2vec.txt που παράγει ο Trainer το καθαρίζει και το προσθέτει στα queries αφού ταυτοποιηθεί η λέξη του query με την πρώτη λέξη των συνωνυμών και αφαιρεθεί.

```
try {
    File myObj = new File( pathname: "w2vec-final.txt");
    Scanner myReader = new Scanner(myObj);
    int i = 0, j = 0;
    String[] list = new String[40];
    while (myReader.hasNextLine()) {
        String data = myReader.nextLine();
        if (data.replaceAll( regex: "[\\W\\W]", replacement: "").equals(""))
            continue;
        list[i] = data.replaceAll( regex: "[\\W\\W]", replacement: "");
        System.out.println(list[i]);
        i++;
    }
    System.out.println("\n\tEnd of synonym list\n");
    myReader.close();

    for (int c = 0; c < 10; c++) {
        queries[c]=queries[c].toLowerCase();
        while (true) {
            if (!(list[j] == null || list[j] == "")) {
                String[] temp = list[j].split( regex: " ", limit: 2);
                if (queries[c].contains(temp[0])) {
                    queries[c] += " " + temp[1];
                    j++;
                } else if (!(queries[c].contains(temp[0]))) {
                    break;
                }
            } else if (list[j] == "") {
                continue;
            } else if (list[j] == null) {
                break;
            }
        }
        queries[c].trim().replaceAll( regex: "+", replacement: " ");
        System.out.println(queries[c]);
    }
    System.out.println("\n\tEnd of queries list and Start of Results log\n");
}
```

Το αρχείο που χρησιμοποιείται παραπάνω είναι το w2vec-final.txt και είναι αυτό που με τα συνωνύμα που περιείχε έβγαλε τα καλύτερα αποτελέσματα και όχι το w2vec που αναφέρθηκε παραπάνω καθώς αφού παράχθηκε w2vec.txt και πέτυχα συνολικά 80 relative retrieved στα πρώτα 50 του κάθε query έφτιαξα 2 νέα αρχεία τον Test και τον Tester οι οποίοι μέτρησαν περίπου πόσες φορές εμφανίστηκαν στο σύνολο των κειμένων οι trimmed από τον English Analyzer όροι στο σύνολο των κειμένων με αποτέλεσμα:

Tester{key='multimod', counter=187} Tester{key='travel', counter=278} Tester{key='servic', counter=3794} Question:1 sum:4,259	Tester{key='mobil', counter=1961} Tester{key='servic', counter=3794} Tester{key='tool', counter=4859} Question:6 sum:10,614
Tester{key='big', counter=865} Tester{key='data', counter=6094} Tester{key='mobil', counter=1961} Question:2 sum:8,920	Tester{key='fragment', counter=337} Tester{key='iot', counter=1363} Tester{key='feder', counter=111} Question:7 sum:1,811
Tester{key='european', counter=25} Tester{key='logist', counter=656} Tester{key='applic', counter=6510} Question:3 sum:7,191	Tester{key='seamless', counter=244} Tester{key='effici', counter=5348} Tester{key='european', counter=25} Tester{key='travel', counter=278} Question:8 sum:5,895
Tester{key='architectur', counter=1126} Tester{key='big', counter=865} Tester{key='data', counter=6094} Tester{key='analyt', counter=1086} Question:4 sum:9,171	Tester{key='cross', counter=3525} Tester{key='domain', counter=1097} Tester{key='orchestr', counter=158} Tester{key='servic', counter=3794} Question:9 sum:8,574
Tester{key='architectur', counter=1126} Tester{key='industri', counter=3736} Tester{key='iot', counter=1363} Question:5 sum:6,225	Tester{key='commun', counter=4085} Tester{key='network', counter=4186} Question:10 sum:8,271

Αυτό με βοήθησε στο να δω πόσα θα συνώνυμα θα χρησιμοποιήσω από τα 15 που είχα διαθέσιμα από το w2vec για κάθε όρο, καθώς ένας όρος με πάρα πολλές ή πάρα πολύ λίγες επαναλήψεις στο σύνολο των κειμένων δεν θα βοηθούσε στην επίτευξη του σκοπού μας καθώς θα επέστρεφε άχρηστα κείμενα τα οποία θα κατέβαζαν τα ποσοστά επιτυχίας των επιθυμητών.

Άρα για όρους όπως το commun= community , network= networks που έχει πάρα πολλές επαναλήψεις ο καθένας 4000+ χρησιμοποίησα μόνο 3 συνώνυμα και το αρχικό query, τα πρώτα 3 που μου επέστρεψε το μοντέλο για την περίπτωση αυτή τα:

```
community [communities, stakeholders, engagement]
networks [network, communication, communications]
```

Αντίστοιχα για τον όρο 'european' που είχε μόνο 25 εμφανίσεις επίσης χρησιμοποίησα μόνο 3 συνώνυμα και το αρχικό query για να μην μπερδευτούν αυτές οι λίγες εμφανίσεις με «κακά» συνώνυμα

```
european [europe, eu, international]
```

Τέλος για ένα όρο όπως το 'big' με «μέτριες» εμφανίσεις χρησιμοποίησα και τα 15 συνώνυμα για να καθορίσουν καλύτερα το πλαίσιο των κειμένων που ψάχναμε, για παράδειγμα :

```
big [analytics, data, data-driven, cloud, intelligence, computing, collection, privacy, mining, bigmedilytics, cloud-based, queries, veracity, unstructured, companies]
```

Σε αυτό το σημείο να τονίσω για να μην παρεξηγηθώ πως δεν επέλεξα ποια συνώνυμα να κρατήσω και ποια όχι, ούτε άλλαξα την σειρά που μου τα «έγραψε» το μοντέλο αλλά έδωσα χειρωνακτικά διαφορετική βαρύτητα «έθεσα βάρη» στα συνώνυμα βάση της εμφανίσεις που είχε ο κάθε όρος και συνολικά το query χωρίς πάντα να υπολογίζω το αρχικό.

Έτσι κατέληξα στο w2vec-final.txt:

```
multimodal [multi-modal, modalities, neuroimaging, imaging, perception, speech, fbi, cognitive, stimulation, tactile, facial, modality, virtual, sensorimotor, fMRI]
travel [passengers, trip, ticketing, door-to-door, destination, booking, travellers, passenger, dora, journey, traffic, seamless, traveller]
services [service, support, management, smes, providers, innovation, users, provision, business, access, infrastructure]

big [analytics, data, data-driven, cloud]
data [information, big, analytics, analysis]
mobility [urban, cities, city]

european [europe, eu, international]
logistics [logistic, transportation, freight, goods, intermodal, last-mile, chain, shipping, planning, supply, providers, cargo, chains, automation]
applications [application, devices, technologies]

architectures [reconfigurable, computing, architecture]
big [analytics, data, data-driven, cloud, intelligence, computing, collection, privacy, mining, bigmedilytics, cloud-based, queries, veracity, unstructured, companies]
data [information, big, analytics, analysis, datasets]
analytics [big, data, real-time, intelligence, algorithms]

architecture [architectures, hardware, programmability, reconfigurable, computing, heterogeneous, scalable, scalability, exascale, low-power]
industrial [industry, industries, partners, sectors, manufacturing, consortium, products, sector, production, academic]
iot [things, internet, smart, cloud, platforms, connected, interoperability, services, ict, end-to-end, security, secure]

mobility-as-a-service [my-trac, car-sharing, last-mile, crowd-sourced, seamless, maas, av-ready, mycorridor, sharing, services, arcadia, city, socialcar, keyless, service]
tools [methods, techniques, approaches, develop, models, tool, provide, analysis, methodologies, framework, developing, modelling, existing]

fragmentation [fragmented, ecosystem, synergies, bilbao, fragclim, coalescence, mass, atacama, ensuf, era-min, observatory, biodiversity, observations, initiatives]
iot [things, internet, smart, cloud, platforms, connected, interoperability]
federation [federated, fed4fire, cloud, testbeds, interoperability, roaming, end-to-end, iot, interoperation, open, orchestration, federations, infrastructure, testbed, deployment]

seamless [interoperability, interoperable, user, end-to-end, users, services, sharing, service, providers, secure, platform, mobile, trusted, user-centric]
efficient [efficiency, solutions, technologies, generation, highly]
european [europe, eu, international, countries, national, innovation]
travelling [travellers, travel, ipco, door-to-door, openings, mobility, cloudfone, nuisance, enclosures, trip, seamless, ionosphere, home, fires, travelers]

cross-domain [interoperability, orchestration, agile, instantiate, 5gex, multi-vendor, iot, sonata, nfv, end-to-end, multi-domain, virtualized, interoperable, middleware, multi-tenancy]
orchestration [end-to-end, virtualization, nfv, virtualised, multi-tenancy, 5g-ready, 5gpagoda, virtualized, multi-operator, programmability, 5g]
services [service, support, management]

community [communities, stakeholders, engagement]
networks [network, communication, communications]
```

Το οποίο περιέχεται στο αρχείο για να επαληθεύσετε τα αποτελέσματα και είναι επιλεγμένο να διαβαστεί από το Searcher για να γίνει η αναζήτηση.

Αν τρέξετε την Main έχω σε σχόλια τον Trainer που παράγει το μοντέλο, και αν αποφασίσετε να τον τρέξετε θα πρέπει να αλλάξετε το αρχείο που διαβάζει ο searcher σε w2vec.txt για να διαβάσει τα συνώνυμα που παράχθηκαν.

Το συμπιεσμένο αρχείο περιέχει σε μορφή png με όνομα Stats-Report.png, τα παρακάτω στατιστικά αρχεία σε περίπτωση που δεν φαίνονται καλά.

Μέρη:		1ο Μέρος Εργασίας			2ο Μέρος Εργασίας			3ο Μέρος Εργασίας		
Query ID	k	avgPre@k	Rel Ret Doc	MAP@k	avgPre@k	Rel Ret Doc	MAP@k	avgPre@k	Rel Ret Doc	MAP@k
Q01	5	0.1698	4		0.1698	4		0.1719	3	
	10	0.3857	8		0.3857	8		0.3537	7	
	15	0.5307	11		0.4785	10		0.4424	9	
	20	0.5749	12	0.5749	0.5167	11	0.5167	0.5097	11	0.5097
	30		14	0.6516		14	0.6133		11	0.5097
	50		15	0.6792		15	0.6351		16	0.6062
Q02	5	0.1389	2		0.0278	1		0	0	
	10	0.1746	3		0.0516	2		0	0	
	15	0.1746	3		0.0516	2		0.006	1	
	20	0.1746	3	0.1746	0.0663	3	0.0663	0.006	1	0.006
	30		3	0.1746		3	0.0663		1	0.006
	50		3	0.1746		3	0.0663		4	0.026
Q03	5	0.2536	4		0.2536	4		0.0833	2	
	10	0.3743	6		0.3743	6		0.1048	3	
	15	0.3743	6		0.3743	6		0.1252	4	
	20	0.4323	8	0.4323	0.4323	8	0.4323	0.1475	5	0.1475
	30		11	0.5078		11	0.5078		6	0.1623
	50		14	0.5745		14	0.5745		10	0.2272
Q04	5	0.0464	2		0.0464	2		0.0357	1	
	10	0.0464	2		0.0464	2		0.0561	2	
	15	0.0464	2		0.0464	2		0.0561	2	
	20	0.0598	3	0.0598	0.0583	3	0.0583	0.0561	2	0.0561
	30		3	0.0598		3	0.0583		2	0.0561
	50		4	0.0688		4	0.0653		6	0.087
Q05	5	0.0896	3		0.0896	3		0.0729	2	
	10	0.0896	3		0.0896	3		0.1042	3	
	15	0.0896	3		0.0896	3		0.1208	4	
	20	0.1207	5	0.1207	0.1207	5	0.1207	0.1208	4	0.1208
	30		9	0.1911		9	0.1911		7	0.0775
	50		13	0.2619		13	0.2619		10	0.2084
Q06	5	0.0263	1		0.0175	1		0.1447	3	
	10	0.0263	1		0.0175	1		0.2978	7	
	15	0.0263	1		0.0175	1		0.3329	8	
	20	0.0263	1	0.0263	0.0175	1	0.0175	0.3329	8	0.3329
	30		2	0.0304			0.0212		9	0.3355
	50		4	0.0391		3	0.025		10	0.369
Q07	5	0.0625	1		0.0625	1		0	0	
	10	0.0764	2		0.0764	2		0.0228	2	
	15	0.0934	3		0.0934	3		0.0399	3	
	20	0.0934	3	0.0934	0.0934	3	0.0934	0.053	4	0.053
	30		9	0.188		9	0.188		6	0.0775
	50		12	0.2383		12	0.238		8	0.0975
Q08	5	0.3571	5		0.3571	5		0.194	4	
	10	0.5714	8		0.5714	8		0.3773	7	
	15	0.6143	9		0.5714	8		0.4828	9	
	20	0.6982	11	0.6982	0.6489	10	0.6489	0.4828	9	0.4828
	30		11	0.6982		12	0.7056		10	0.5139
	50		11	0.6982		12	0.7056		13	0.579
Q09	5	0.131	3		0.0714	2		0.2381	5	
	10	0.15	4		0.0714	2		0.4286	9	
	15	0.1683	5		0.0833	3		0.5026	11	
	20	0.1683	5	0.1683	0.0934	4	0.0934	0.5706	13	0.5706
	30		7	0.1936		4	0.0934		16	0.6611
	50		9	0.2126		5	0.0981		18	0.7044
Q10	5	0.05	1		0	0		0	0	
	10	0.075	2		0.0125	1		0	0	
	15	0.075	2		0.0125	1		0	0	
	20	0.075	2	0.075	0.0125	1	0.0125	0	0	0
	30		3	0.0853		1	0.0125		0	0
	50		3	0.0853		2	0.0181		4	0.0241
All Queries	5	0.1325	26		0.1096	23		0.0941	20	
	10	0.197	39		0.1697	35		0.1745	40	
	15	0.2193	45		0.1819	39		0.2109	51	
	20	0.2424	53	0.2424	0.206	49	0.206	0.2279	57	0.2279
	30		72	0.278		68	0.2457		68	0.2508
	50		88	0.3033		83	0.2688		99	0.2929