In vocabulary seed generation for out of vocabulary seeds in seed guided topic modelling

Tiancheng (Jackson) Xia

I preface by first explaining that the final testing included manual collection of text from websites due to IP ban from google for misusing their service with my web crawler bot, whilst this does potentially mean that the data is now inherently biased, this would in fact better represent a practical use case as we should be using select databases and extracting only the text rather than everything on the page of any website searchable through google.

Table of Contents

Abstract

The discovery of topics from large text corpora has been used for a wide range of purposes and serves to enable other natural language processing tasks with its ability to quickly gain information on raw textual data and grouping them. Many existing topic models that achieve this does so through an unsupervised approach, which depending on how similar the topics are, may fail due to its inability to incorporate user guidance. Whilst there exist seeded approaches, that uses words for topics as guidance, they fail to perform if the words do not appear in the corpus. In this paper, instead of introducing a new topic modelling algorithm that will utilise out of vocabulary seeds, I aim to prove a concept of generating seed words from out of vocabulary seeds through a different but larger corpus which, when applied properly, would solve the problem universally. The experiments demonstrate the viability of the seed generation with quality data, however, it failed to methodically prove its full implementation during the given timeframe. The future of this concept, however, is promising as shown with further experimentation, and the suggestions for a large scale implementation has been proposed for future works.

Introduction

With the advancements in technology in recent years, the amount of data in the form of textual information has simply been too much for manual collection and labelling. This, coupled with the ease of both accessing and publishing data whether through starting their own website or a post on their social media, further makes this a problem for users searching for texts within massive text corpora. Topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been efficient in its discovery of informative topics from massive text corpora and was one of the earlier algorithms that allowed management and analysis of large text corpus. LDA achieves this by generating random documents by choosing a mixture of topics and its corresponding words for that document, which we can then use to calculate the topic-word and document-topic distributions that closely matches to the documents from the text corpus (Rinfret, 2019). The reason why the output are distributions is because documents don't necessarily have just one topic, meaning that they have a distribution and similarly with topics as for example, both "galaxy" and "stars" belong to the topic "astrophysics" with varying probability. LDA however, due to its unsupervised nature often leads to topics that are simply far from user interests, particularly if the text corpus has not been cleaned up properly. For example, if the text corpus was automatically generated from webpages and contained words such as "checkout" and "sale", when the search query was on computer parts, due to stores selling these computer parts then these words could potentially form a topic instead of computer parts when using LDA. Given these constraints, semi supervised approaches have been introduced and adopted due to their ability to incorporate lexical priors or seeds to refine the process and generate topics like the seeds provided

(Jagarlamudi et al., 2012). This however brings us to the next problem, whereby the seeds are required to appear in the text corpus. Currently, it is standard practice to check and verify input seeds appears, and remove the out of vocabulary seeds, which in a highly automated process could result in the lack of seeds for a particular topic. The problem is largely unsolved and the only publicly available and recently proposed way of incorporating out of vocabulary seeds from SEETOPIC through semantic analysis of seeds to get an understanding of the meaning of the seed through the power of Pretrained Language Models (PLM) (Zhang et al., 2022), this report aims to show a proof of concept through a different approach.

Due to the nature of the proposed approach, instead of addressing this issue with a concrete algorithm, a general framework for the algorithm will be proposed due to its high customisability. The framework is quite simple, which is to generate in vocabulary seeds from out of vocabulary seeds by using a seeded topic model to find words to replace the out of vocabulary seed from texts gathered from the internet. For the purposes of this project, due to computing and time constraints google search and web scraping texts were automated for the initial testing phases before manual collection of text online for the presented data due to terms of service requirements. Furthermore, GuidedLDA package was used as the primary seeded topic model for its ease of use. Links to these will be provided in the source code for the project under the readme file.

Exploratory analysis

Prior to running the program on a dataset, an initial quick test was done to see the effectiveness of the set up as follows:

1. Google search each out of vocabulary seed *
    a. Select number of pages of search results for each seed
    b. Collect all URL
    c. Filter unwanted URL
2. Collect all text from each URL
3. Pre-process the text with stemming and lemmatization
4. Use GuidedLDA on all text with the out of vocabulary seeds as seed
5. Record the top words for each topic

* **DO NOT RUN THIS PART OF THE CODE WITHOUT INCREASING SLEEP TIME OR GOOGLE WILL BLACKLIST YOUR IP**

This code was then run to test two different approaches in mind: multiple different searches for each topic such as Nasa, stars, moon etc for space and only one search but more pages of results or more texts for each topic. This was done for a few reasons, one of which was to check whether multiple different google search queries would be able generate similar enough keywords. This concern is because the text corpus for the seed generation for a particular topic would no longer be a singular topic, but rather a collection of subtopics for that parent topic, potentially causing the topic-word distribution to bias toward the subtopics searched and/or biasing towards the overlap of the subtopics which could be very generic. For example, the aim is to generate documents around space via google search, and the keywords such as Nasa, stars, moon have been used, whilst one may safely assume that these keywords are a subset of space, the act of grouping them together can potentially cause issues as the topic-word distributions for these keywords is not equivalent to the topic-word distributions of space. However, based on Figure 1, GuidedLDA performs decently well in both the multiple seed search and single seed search with a few caveats and potential features which will be addressed.

Results

Topic	Multiple seed searches representative words	Single seed search representative words
Warfare	nuclear, weapon, war, state, world, warfar, unit, use, bomb, militari, countri, intern, forc, arm, soviet, power, develop, atom, nation, attack, union, secur, effect, chemic, american	war, warfar, militari, state, san, nuclear, duti, modern, antonio, edit, weapon, oper, forc, retriev, histori, world, unit, conflict, new, battl, citi, includ, american, arm, texa
space	space, nasa, imag, new, launch, monday, earth, scienc, station, mission, russia, time, orbit, moon, juli, satellit, year, russian, telescop, univers, war, open, intern, star, astronaut	planet, moon, space, earth, solar, retriev, orbit, nasa, scienc, doi, lunar, origin, sun, object, star, new, archiv, bibcod, jupit, telescop, webb, planetari, time, univers, natur
medicine	health, medicin, diseas, medic, care, covid, research, cancer, studi, retriev, includ, prevent, patient, organ, scienc, edit, treatment, archiv, practic, develop, clinic, surgeri, origin, physician, inform	medicin, medic, health, research, care, educ, patient, clinic, school, scienc, student, studi, program, diseas, practic, univers, physician, cancer, inform, faculti, surgeri, drug, treatment, year, includ
Travel and transport	tourism, travel, outdoor, new, activ, servic, sport, event, manag, india, read, day, share, transport, news, busi, tour, place, cultur, home, park, recreat, shop, experi, visit	travel, servic, transport, busi, manag, inform, data, use, industri, time, polici, share, support, new, provid, compani, work, search, plan, oper, custom, contact, access, world, experi
Sports and outdoors	game, team, say, season, fan, player, vaccin, leagu, test, covid, play, event, week, sport, schedul, posit, year, day, new, allow, start, state, athlet, time, open	sport, outdoor, republ, academi, accessori, bike, sale, gear, fish, shop, news, equip, park, deal, new, star, store, view, game, camp, cloth, shoe, und, email, run

Figure 1, table of seeds generated with automated google searches.

The seeds are the way they are due to lemmatization (the very basic form of the word) and will be kept this way for the entirety of this paper

Analysis of results

The data extraction and pre-processing are sub-par and results are subject to noise.

Based on the results above, words like "email" and "edit" have appeared due to the complex nature of scraping text from the internet through purely URL alone. As different websites have different structures, all text and information were extracted instead and that resulted in some noise in the dataset. This will be addressed by manually collecting the text from each website instead for the final test.

Google search and PageRank.

The google search and the underlying PageRank algorithm is secretive and prioritises certain websites over others such as Wikipedia and YouTube and news websites when a certain event occurs. This will introduce randomness and unwanted bias towards the results and make certain results unable to be replicated. For example we can see in "Sports and Outdoors", one of the seeds was "upcoming outdoor sporting events" which then resulted in "covid" and "test" both becoming top 10 related words due to recent world events, which could prove to be quite useful at times compared to PLMs as since they're pretrained, it might fail to recognise semantic links between new phrases if they are not updated regularly. This caveat/feature cannot be reasonably avoided using the current setup, however since the focus of this project is a proof of concept of using a larger database like the internet to generate similar words, this is something that can be either utilised or avoided in a larger project.

Lack of semantics and phrases.

From the data in the table above, it becomes quite clear that GuidedLDA has areas that require improvements, as seen with the "test" and "covid". Semantically, it doesn't make much sense to include covid testing with sports and outdoors under normal circumstances, coupled with the noise that comes from recent events that PageRank will recommend first, this could become common occurrence. Furthermore, a lot of meaning is lost during the pre-processing stage, as the phrases are broken down into individual words such as "covid test" and potentially "Webb telescope images". This would decrease the effectiveness of the model if there were two similar and overlapping topics such as "world war I" and "world war II". This, however, is a constraint that is particular GuidedLDA, and a proposed solution will be proposed in the discussion section but for the sakes of this project, topics with similar names will be avoided.

Power of seeds as topics of a dataset.

It is quite clear that if the topics that the user is interested in have little to no semantic overlap, seeded datasets can generate topics with words that bias towards the dataset required. For example, in the travel and transport section seeds like "tourism" and "overseas travel and transport" were used and thus resulted in words like "culture", "recreation" and "tour", which is expected as "the quality of the seed topics is determined by the discriminative power of the seed words" (Jagarlamudi et al., 2012).

Dataset

The dataset this algorithm will be using is the Kaggle's "Topic Modeling for Research Articles" dataset posted by Abishek Sudarshan as it is one of the few labelled topic modelling datasets that's free to access and will be slightly modified to accommodate for the inability to utilise automated text scraping through google due to temporary IP bans. The dataset will be broken down and only 4 tags will be kept, and will be "Astrophysics of Galaxies", "Fluid Dynamics", "Robotics" and "Superconductivity".

First experiment

The effectiveness of the model will be evaluated in different stages. Firstly, we will evaluate the relevancy of the top representative words of each topic. Then we will run the setup against both GuidedLDA (with topic names as seeds and removed if doesn't appear) and LDA to test the efficiency of the set up. Since our aim is to see if an external corpus could generate useful in vocabulary seed words from out of vocabulary seeds, it is essential that we keep the underlying algorithm the same and so if we were to change the setup from seeded LDA to CatE then we would also need to use CatE as a baseline for a fair evaluation.

Algorithm

```
from $D1$, the corpus with out of vocabulary seeds, collect text corpus $D2$ from web pages or other databases using either out of vocabulary seeds or topic names from $D1$ as search query C
for $d$ in $D2$ do
    preprocess(d) // stemming, lemmatise, remove punctuation and \n
GuidedLDA(D2,C)
for $t \leftarrow 1\ to\ |T|$ do
    select top n most representative words of topic $t$ and save as seed set $S$
GuidedLDA(D1,S)
```

Results

The results of the seed generation are as follows:

Topics	Generated seed words
Astrophysics of Galaxies	galaxi, star, univers, observ, mass, way, cluster, gas, studi, disk, format, larg, matter, stellar, black, milki, hole, astrophys, evolut, dark, project, form, structur, activ, astronom, light, spiral, time, center, telescop, process, understand, distribut, measur, sun, radio, physic, survey, research, year, like, new, data, gravit, contain, object, cloud, scale, includ, know
Fluid Dynamics	flow, fluid, model, simul, time, dynam, equat, method, comput, turbul, general, forc, differ, solut, chang, number, learn, numer, scale, approach, problem, accuraci, result, cfd, train, veloc, pressur, energi, point, use, volum, physic, field, step, reynold, solv, base, calcul, analysi, requir, direct, high, condit, refer, predict, resolut, appli, continu, effect, accur
Robotics	robot, human, control, develop, task, design, industri, engin, use, research, build, machin, power, perform, program, exampl, work, like, environ, system, includ, scienc, technolog, oper, autonom, team, learn, competit, mechan, new, sensor, type, interact, help, inform, need, actuat, intellig, product, creat, requir, process, motor, world, advanc, construct, walk, applic, data, concept
Superconductivity	temperatur, superconductor, magnet, field, materi, superconduct, high, current, energi, electron, state, theori, critic, electr, transit, type, resist, properti, pair, call, effect, low, normal, discov, zero, heat, know, element, metal, cooper, pressur, electromagnet, quantum, phase, appli, physicist, discoveri, applic, sampl, explain, measur, bcs, meissner, higher, order, physic, thermal, theoret, generat, gap

Figure 2, table of generated seed words from text manually collected from webpages of google search results

Of the 50 total seed words on Astrophysics of Galaxies (as the judges are more familiar with this topics from high school physics) in Figure 2, based on majority vote between 5 judges, we've found that 28% of the seeds generated are unrelated (green) and 14% of the seeds generated would only correlate to the topic with context (blue).

From the results of the 2000 training and 427 test cases, we can see quite clearly that the topic-word distributions for the 3 algorithms have converged to a similar state, where the top 10 representative words are mostly the same with a few words having swapped order of significance. Furthermore, of the 427 test cases, GuidedLDA with a seedless topic, LDA and GuidedLDA with generated seeds have had similar accuracy with 97.4% with low variance depending on the random state of the algorithm.

Topics\algorithm	GuidedLDA with a seedless topic	GuidedLDA with generated seeds	Normal LDA
Astrophysics of Galaxies	galaxi, star, mass, observ, cluster, gas, stellar, format, model, emiss, line, halo, high, sourc, densiti, data, sim, redshift, sampl, measur, low, time, region, help, field, present, form, detect, result, survey, larg, dust, veloc, studi, luminos, simul, relat, disk, galact, massiv, radio, ray, distribut, alpha, metal, agn, differ, evolut, consist, scale	galaxi, star, mass, observ, cluster, gas, stellar, format, model, emiss, line, halo, high, sourc, densiti, measur, sim, data, redshift, sampl, time, present, region, low, detect, help, veloc, form, survey, field, dust, larg, studi, luminos, result, galact, disk, massiv, ray, relat, radio, alpha, evolut, simul, agn, metal, distribut, properti, scale, consist	galaxi, star, mass, observ, cluster, gas, stellar, format, model, emiss, high, line, halo, densiti, sourc, sim, data, redshift, sampl, measur, time, low, region, form, present, result, detect, survey, dust, veloc, studi, larg, luminos, help, field, disk, galact, relat, massiv, radio, evolut, ray, simul, distribut, alpha, agn, metal, scale, properti, differ
Fluid Dynamics	flow, model, equat, simul, number, turbul, fluid, dynam, numer, effect, veloc, energi, wave, time, particl, result, solut, field, studi, surfac, method, scale, non, boundari, forc, differ, condit, case, pressur, paramet, order, analysi, direct, help, depend, mode, magnet, linear, interact, stabil, larg, present, mechan, state, phase, structur, dimension, nonlinear, deriv, show	flow, model, equat, simul, turbul, fluid, number, result, dynam, numer, energi, particl, wave, veloc, effect, field, solut, scale, time, surfac, studi, method, non, forc, condit, help, boundari, pressur, direct, order, case, magnet, larg, analysi, paramet, linear, stabil, depend, differ, interact, mechan, distribut, dimension, mode, small, increas, state, present, nonlinear, structur	flow, model, equat, number, simul, turbul, fluid, dynam, numer, result, veloc, particl, energi, effect, field, wave, surfac, solut, studi, scale, method, time, non, help, forc, magnet, condit, pressur, direct, boundari, case, stabil, differ, order, analysi, depend, larg, linear, mode, state, mechan, paramet, present, nonlinear, structur, interact, dimension, deriv, increas, instabl
Robotics	robot, method, learn, model, control, propos, base, object, task, algorithm, help, perform, problem, idea, human, environ, paper, present, time, result, data, simul, demonstr, real, network, train, optim, imag, plan, use, high, motion, state, dynam, differ, trajectori, comput, map, provid, experi, effici, framework, polici, generat, challeng, design, vehicl, sensor, system, predict	robot, method, learn, model, control, propos, base, task, object, algorithm, help, perform, problem, environ, human, idea, paper, present, simul, data, time, result, demonstr, network, train, real, optim, imag, plan, use, high, differ, state, dynam, motion, map, trajectori, experi, generat, provid, polici, framework, predict, challeng, comput, effici, design, vehicl, sensor, system	robot, method, learn, model, control, propos, base, object, task, algorithm, help, perform, problem, human, idea, environ, paper, present, time, simul, data, result, demonstr, network, real, train, optim, plan, use, imag, state, high, differ, dynam, motion, map, trajectori, provid, framework, comput, effici, predict, polici, experi, generat, challeng, design, vehicl, sensor, system
Superconductivity	magnet, state, superconduct, temperatur, phase, electron, spin, field, superconductor, order, transit, pair, quantum, high, coupl, structur, effect, result, band, gap, symmetri, studi, dope, interact, wave, critic, observ, topolog, strong, energi, measur, model, fermi, layer, lattic, densiti, depend, singl, orbit, charg, base, pressur, scatter, experiment, metal, fluctuat, crystal, low, investig, help	magnet, state, superconduct, temperatur, phase, electron, spin, order, field, superconductor, transit, pair, quantum, high, coupl, structur, effect, interact, band, result, gap, critic, symmetri, dope, observ, studi, topolog, wave, strong, energi, model, measur, fermi, layer, depend, lattic, singl, orbit, charg, pressur, base, experiment, densiti, scatter, fluctuat, crystal, metal, paramet, low, investig	magnet, superconduct, temperatur, state, phase, electron, spin, field, order, superconductor, transit, pair, quantum, high, coupl, interact, band, result, effect, structur, gap, dope, wave, symmetri, studi, observ, topolog, critic, strong, energi, measur, fermi, layer, model, depend, lattic, singl, orbit, charg, base, densiti, metal, fluctuat, pressur, scatter, crystal, experiment, low, paramet, system

Figure 3 table of top words from topic-word distributions of GuidedLDA, GuidedLDA with generated seeds and LDA

Discussion

Since the semantically unrelated words are quite general, these words have little effect on the topic-word distributions when used as seeds due to their low discriminative power (Jagarlamudi et al., 2012). The results in Figure 2 along with Figure 1, proves the viability of the first component of the concept, which is the generation of semantically similar seeds from out of vocabulary seeds. It is worthy to note that only 57 total web page search results were used due to labour intensive nature of manual collection of these texts and data cleaning to make sure words like "edit", "advertisements" and equations are removed. Furthermore, since most of the unrelated words are populated towards the end, it is highly likely that these seeds have resulted due to the high variance associated with a low amount of training data or small corpus.

Based on the results in Figure 3 and the similar accuracy achieved from all the cases, it is quite clear that the three algorithms have converged to an almost equivalent topic-word and document-topic distributions, likely since the text corpus for both the testing and training phase is too different for each topic. This unfortunately does not prove the core aim of the project and repeating the experiment for a new and more difficult corpus is not possible in the current timeframe, hence we look for a new way to prove our concept.

Second experiment

With the results from the first experiment, we change our approach by lowering the amount of training data to simulate scenarios where finding a satisfactory topic-word distribution becomes really hard but a well seeded GuidedLDA will perform well in comparison. With a small training data set, the result is subject to high variance due to the inherent noise associated with each individual document and so a solution to that will be to take the median of 5 runs of the algorithm for each training data size with each run using a different random state. Since the dataset is labelled, we may compare the performance of the three algorithms under these scenarios.

Results

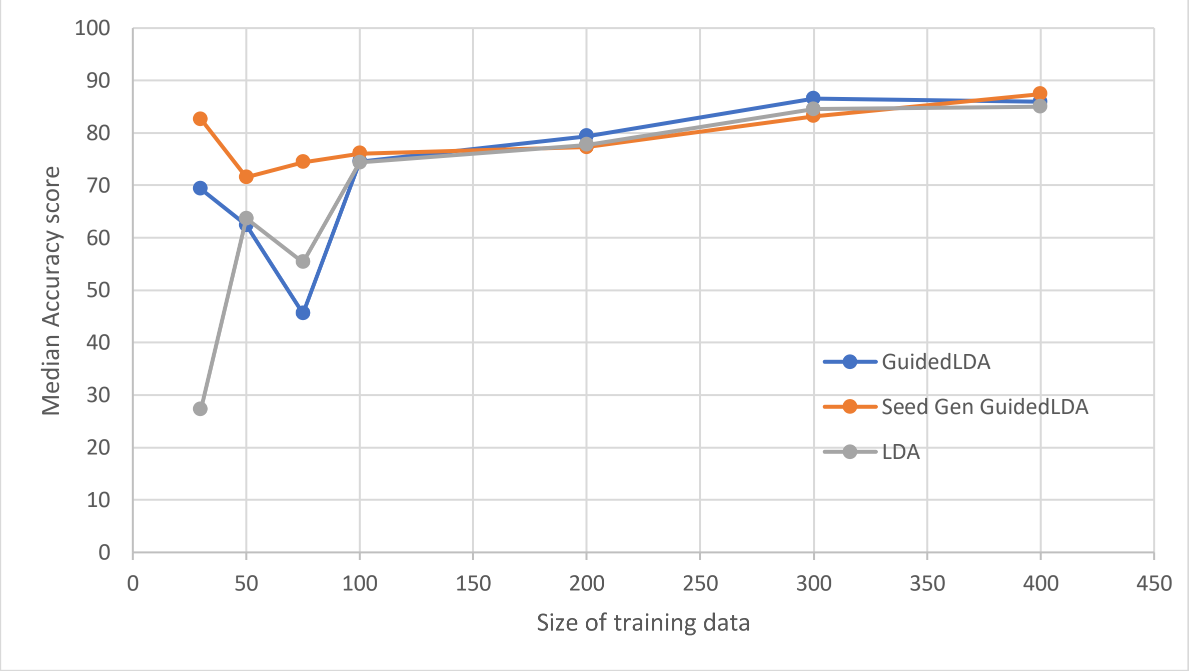

Figure 4, median accuracy score of 5 simulations with random "random_state" based on size of training data

The results of the median of 5 for the three algorithms on different training data is a good representation of also the discarded simulations, particularly when the training data is < 100. The difference between min and max accuracy scores of the seed generated GuidedLDA were all within 10% except for training data size 30. The max of LDA and GuidedLDA simulations were not greater than 5% higher than max of seed generated GuidedLDA, however the min of LDA and GuidedLDA were extremely low due to their inability to capture a correct topic-word distribution which is very clear in Figure 5. In GuidedLDA with a seedless topic, we see that there is a clear misclassification where words semantically regarding optics and astrophysics like "emission", "infrared", "spectra", "quasar" and others are in the fluid dynamics section. Similarly, for LDA, under fluid dynamics section, we have the semantically galaxy words and fluid dynamic words mislabelled into other sections like robotics. GuidedLDA with generated seeds, however, has almost all the right topic-word distributions, with all its top 10 words related semantically to its topic or in no topics likely due to overfitting to documents. Furthermore, from the collection of all results for GuidedLDA with generated seeds apart from size 30 training data, the algorithm had relatively low variance when compared to its counterparts, and even in its lowest accuracy runs for these dataset sizes, at no point did any top words for topics contain words that was clearly from another topic.

Topics\algorithm	GuidedLDA	GuidedLDA with generated seeds	Normal LDA
Astrophysics of Galaxies	galaxi, mass, star, observ, model, present, simul, data, format, region, gas, stellar, sampl, cluster, low, redshift, time, compar, show, field, deriv, relat, help, imag, sim, result, halo, detect, signific, radio, line, approxim, effect, metal, high, number, consist, scale, ray, evolut, measur, includ, odot, extend, differ, densiti, rang, feedback, size, larg	galaxi, mass, observ, star, format, gas, stellar, data, cluster, redshift, region, emiss, ngc, present, low, disk, radio, halo, line, dwarf, metal, ray, infrar, evolut, deriv, sim, odot, compar, sampl, simul, dust, suggest, age, densiti, imag, sourc, resolut, hole, previous, includ, form, new, spectra, field, black, bar, object, central, cloud, feedback	mass, galaxi, star, observ, model, region, metal, low, format, stellar, time, sampl, object, predict, redshift, effect, simul, result, deriv, sim, relat, scale, detect, line, help, high, differ, includ, consist, measur, show, approxim, signific, evolut, odot, feedback, present, larg, function, sourc, mean, uncertainti, size, small, estim, black, hole, locat, larger, self
Fluid Dynamics	observ, emiss, ngc, system, disk, studi, properti, model, natur, dwarf, previous, infrar, distribut, veloc, new, predict, recent, dust, paramet, base, interact, associ, age, motion, magnitud, gmas, larger, near, lead, activ, evid, post, spectra, type, limit, matter, alma, strong, dynam, optic, photometri, scatter, hydrodynam, compon, quasar, agreement, rotat, similar, decreas, long	model, result, time, help, effect, high, system, studi, larg, differ, predict, scale, measur, dynam, veloc, base, observ, relat, rang, show, signific, approxim, natur, process, local, consist, near, effici, edg, number, paper, simul, paramet, distribut, experiment, recent, physic, function, multi, like, depend, thermal, use, magnitud, strong, uncertainti, select, mean, addit, sampl	gas, cluster, data, observ, veloc, ngc, emiss, disk, halo, radio, galaxi, dwarf, ray, infrar, new, field, previous, dust, model, age, imag, present, compar, spectra, band, alma, optic, resolut, like, photometri, obtain, near, magnitud, power, survey, reproduc, shock, frequenc, gyr, quasar, agreement, type, use, select, ghz, show, lambda, distanc, proxim, bump
Robotics	method, robot, learn, problem, control, propos, model, optim, framework, solut, human, new, inform, help, demonstr, polici, network, feedback, train, time, algorithm, use, comput, behavior, agent, object, task, base, general, generat, signal, result, reinforc, paper, visual, featur, dynam, trajectori, bodi, real, novel, provid, track, import, scheme, order, complex, process, predict, grasp	robot, method, problem, learn, control, propos, optim, framework, human, new, inform, feedback, network, polici, solut, demonstr, object, model, algorithm, train, comput, behavior, agent, task, generat, reinforc, featur, provid, signal, visual, real, novel, trajectori, present, detect, general, interact, environ, complex, grasp, scheme, abl, prior, data, enabl, import, idea, scene, state, perform	method, robot, control, learn, problem, propos, optim, framework, new, solut, human, base, inform, model, result, polici, network, time, featur, algorithm, train, comput, feedback, behavior, agent, use, help, generat, task, paper, demonstr, import, visual, effici, signal, reinforc, trajectori, bodi, real, novel, process, main, requir, track, scheme, provid, prior, dynam, complex, general
Superconductivity	field, structur, magnet, state, superconduct, electron, pair, result, wave, high, spin, transit, pressur, phase, conduct, superconductor, droplet, physic, edg, thermal, turbul, metal, band, correl, direct, effect, mode, flame, suggest, order, measur, drive, temperatur, symmetri, cuprat, fluid, ball, gap, neutron, numer, flow, control, scale, process, presenc, bubbl, analyz, larg, discuss, moder	field, magnet, state, superconduct, electron, structur, pair, wave, order, transit, conduct, spin, core, pressur, properti, phase, associ, high, superconductor, droplet, gmas, turbul, metal, band, correl, temperatur, mode, scatter, form, symmetri, low, mathrm, fluid, neutron, mechan, cuprat, flow, gap, control, suggest, theori, damp, iron, forc, potenti, chiral, term, classic, understand, sdw	field, structur, magnet, system, state, superconduct, properti, high, electron, studi, pair, natur, order, suggest, phase, wave, number, densiti, transit, spin, conduct, physic, core, thermal, pressur, temperatur, addit, investig, superconductor, droplet, strong, motion, associ, local, dynam, interact, turbul, larg, correl, gmas, discuss, activ, result, direct, edg, distribut, drive, recent, post, paramet

Figure 5, table of top representative words for the median of the 5 runs on a training data size of 75.

Discussion

Based on the results of our second experiment it becomes quite clear that the algorithm greatly reduces the variance of different random states and improves the topic models. This was not surprising however as the inclusion of a high amount of mostly quality seeds generated from either in vocab or out of vocabulary seeds in each topic prevents the issues pointed out in the results section for Figure 5 where we have words from one topic spread into two topics. Based on that scenario we would have the topic that was squeezed out to have their words spread into other topics which would decrease the accuracy of the topic model.

Whilst the model shows to be promising, the methodology is unable to prove the concept fully due to the artificial nature of the second experiment's results on the aim. This, however, does not discount the benefits with the algorithm, as with the same dataset, generated seed GuidedLDA had consistent good performance in relation to its basic counterparts and behaves similarly to a well seeded GuidedLDA in the artificial scenario. Since GuidedLDA only had seeds for 3 of the 4 topics (with topic names as seeds) it was only able to perform like LDA. Generated seed GuidedLDA however, did prove that regardless of in vocab or out of vocabulary seeds, if they're in vocabulary to a significantly large and trustworthy database, semantically similar seeds can be generated with ranging levels of quality depending on the quality of the database. With these results, the future of this generic algorithm looks promising as it looks to solve out of vocabulary seeds for all seed guided topic modelling algorithms, with the direction of future works being:

- **The use of other algorithms.** Originally planned to make an improvement on the SEETOPIC framework with this algorithm, however due to their late release of their source code that was simply not possible with the deadline. This should also fix the issue of not having any phrases and maintaining semantic meaning of common phrases.
- **Better source of text for out of vocabulary seed.** Whilst the internet is a good place that should cover all possible seeds, the quality of text acquired from websites are not made equal (ads, commercial websites, company about us, list of scholarly articles for download but not the text themselves). This was only avoided by manually grabbing text for the final testing phase. Predetermined sources like Wikipedia and Britannica for seed generation of out of vocabulary seeds also has its benefits of being able to write script that will extract text better than the extraction of all data from webpage and stripping unrelated information such as format related data.

Conclusion

In this report, we have explored the concept of the generation of in vocabulary seeds from out of vocabulary seeds for topic modelling and have seen benefits in certain use case scenarios. From our limited experimental data, we were only able to conclude that the generation of seeds from topic-word distributions from internet generated text corpus is successful depending on the quality of text generated from the internet. Although the second experimental data is promising on the generated seeds effect in its replacement of out of vocabulary seeds, the benefits of these seeds remain to be formally proven and would require future work such as the use of other algorithms and a better source of texts to maximise the potential of the algorithm.

Reference

Rinfret, J. P. 2019. Latent Dirichlet Allocation https://medium.com/swlh/latent-dirichlet-allocation-lda-eff969bda284

Sudarshan, A. 2022. Topic Modeling for Research Articles https://www.kaggle.com/datasets/abisheksudarshan/topic-modeling-for-research-articles

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. Journal of Machine Learning Research

Zhang, Y, Meng, Y, Wang, X, Wang, S & Han, J 2022, 'Seed-Guided Topic Discovery with Out-of-Vocabulary Seeds'.

Li, C, Chen, S, Xing, J, Sun, A & Ma, Z 2019, 'Seed-Guided Topic Model for Document Filtering and Classification', ACM transactions on information systems, vol. 37, no. 1, pp. 1–37.

Jagarlamudi, J., Daumé III, H., and Udupa, R. 2012. Incorporating Lexical Priors into Topic Models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 204–213, Avignon, France. Association for Computational Linguistics.