70 th International Astronautical Congress (IAC), Washington D.C., United States, 21-25 October 2019.

IAC-19,D5,2,7,x51013

The automatic categorisation of space mission requirements for the Design Engineering Assistant

Audrey Berquand [a]∗ , Iain McDonald [a] , Annalisa Riccardi [a], Yashar Moshfeghi [b]

[a] *Intelligent Computational Engineering Lab, Mechanical and Aerospace Department, University of Strathclyde, 75 Montrose St., G11XQ Glasgow, UK*,
audrey.berquand@strath.ac.uk, iain.mcdonald.2015@uni.strath.ac.uk, annalisa.riccardi@strath.ac.uk

[b] *Computer And Information Sciences Department, University of Strathclyde, 75 Montrose St., G11XQ Glasgow, UK*, yashar.moshfeghi@strath.ac.uk

∗ Corresponding Author

Abstract

To enhance Knowledge Reuse in the field of space mission design, the implementation of Information Retrieval (IR) is key. Topic Modeling (TM) is used to identify, learn and extract topics from a corpus of documents, and can therefore support several IR tasks such as categorisation. This study relies on a common TM method, Latent Dirichlet Allocation (LDA), a probability-based approach. An extensive Wikipedia-based corpus focused on space mission design is collected, parsed, preprocessed, and used to train a general 'Space Mission Design' LDA model. The LDA model is optimised based on the perplexity measure for a range of topics numbers. The topics dictionaries of the retained model are labelled by human annotators, with labels corresponding to spacecraft subsystems. The performances of the general model are evaluated against a set of space mission requirements with a categorisation task. The general model is then used as a base to generate specific LDA models focused on one topic, or spacecraft subsystem. The general LDA model developed in this study proves to be a solid base for the generation of focused LDA models, yielding very high accuracy scores and Mean Reciprocal Ranking.Finally, a semi-supervised LDA model, fed with lexical priors is trained, leading to improved performances of a general model.

Keywords: Topic Modeling, LDA, Machine Learning, categorisation, mission requirements, virtual assistant

Acronyms

AOCS Attitude and Orbit Control System
ECSS European Cooperation for Space Standardisation
ESA European Space Agency
GNC Guidance, Navigation and Control
IR Information Retrieval
LDA Latent Dirichlet Allocation
NLP Natural Language Processing
MRR Mean Reciprocal Ranking
OBDH On-board Data Handling
TF-IDF Term Frequency-Inverse Document Frequency
TM Topic Modeling

1. INTRODUCTION

Experts involved in the early stages of space mission design can spend from 25 to 50% of their work time searching for information [1]. Knowledge Reuse is key to kick-starting a study as experts look into previous similar missions' reports, get an estimate of correct value ranges and validated architectures. Information Retrieval (IR) methods should allow experts to access information more quickly and efficiently. To tackle this issue, this study considers the implementation of Topic Modeling (TM), an unsupervised Machine Learning method used to identify, learn and extract topics from a set of documents. TM supports several IR tasks such as categorisation or Question&Answer. This study presents the first application of a common TM method, Latent Dirichlet Allocation (LDA) [2], to a space mission design corpus. Due to the unavailability of an open-source space mission design corpus, a Wikipedia-based corpus was created to train the LDA model. The trained model dictionaries were labelled by human annotators, leading to the development of the first 'Space Mission Design' LDA model. A categorisation task was used to evaluate the model performances, relying on a corpus of space mission requirements. The LDA method is detailed in Section 2. The corpora used to train the model and to evaluate it are both presented in Section 3. The methodologies to optimise the LDA model and to categorise requirements are described in Section 4. Finally, results obtained with the retained general LDA model, the specific and semi-supervised models are summarised in Section 5. The complete code and corpora are available at https://github.com/strath-ace/smart-nlp.

2. BACKGROUND

2.1 *Unsupervised LDA*

LDA was introduced in 2003 by Blei, Ng and Jordan in [2] as a generative probabilistic model for discrete data collections. TM assumes that a document is a mixture of topics, and an LDA model represents a corpus of documents as a distribution probability over latent (hidden) topic. Similar documents should, therefore, have similar topics distribution. Each latent topics is described by a dictionary, a sorted list of words with their probability to belong to the latent topic. For instance, a *Propulsion* topic's dictionary could include the words *thruster*, *engine* or *propellant*. LDA has been commonly used for IR tasks, such as collaborative filtering and recommender systems [9] or trends forecasting [12, 10]. In the Space field, TM was applied to identify trends in NASA Space Systems Problems reports [8].

Within an LDA model, each document is a probability distribution over topics, and each topic is a probability distribution over words. Hence, the probability distribution of topics T among a corpus of documents can be defined as [9]:

$$p(M|\alpha,\beta) = \int p(\theta|\alpha)(\prod_{i=1}^{N}\sum_{z=1}^{T} p(z|\theta)p(w_i|\beta_z)) \qquad (1)$$

where M is a document composed of N words w_i, z is a topic from a set of latent topics T, $p(z|\theta)$ is a multinomial distribution given by θ and followed by topic z, $p(w_i|\beta_z)$ is the probability that word w_i belongs to topic z given by β_z. β and α are the Dirichlet distribution parameters, θ follows the hyper-parameter α. Figure 1 displays the equivalent graphical representation of a LDA model as presented in [2].

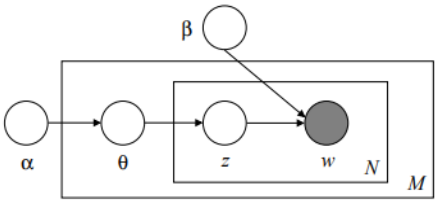

Figure 1: LDA model graphical representation

2.2 *Semi-supervised LDA*

The initial probability distribution of a word to belong to a topic, $p(w_i|\beta_z)$, is randomly set at the start of the modeling process. In the case of the semi-supervised LDA, the probability of certain words to belong to a topic can be increased at the start of the process to influence the composition of the topic dictionaries. The concept of inputting lexical priors, or seed words, into the model is presented in [7]. In the case of the study presented in this paper, the sought-after topics are known, they correspond to spacecraft subsystems. In the Gensim Python library [11] used to train the model, η, a matrix representing for each topic, the probability of each word to belong to it, can be provided to the model to impose the asymmetric priors over the word distribution. The seed words probabilities of each topic is set to 0.95 while the remaining words probabilities are set to 0.

3. CORPORA

This study relies on two corpora: a corpus to train the LDA model, based on Wikipedia pages, and a corpus of requirements for the categorisation application.

3.1 *LDA Model Corpus*

The only observable parameters available to the LDA model are the words contained in the corpus and their frequency. The corpus content needs to be carefully selected to avoid off-topics. In this study, the corpus used to train the LDA models is based on Wikipedia freely available data. The Wikipedia page on Spacecraft Design (https://en.wikipedia.org/wiki/Spacecraft_design) was used as a starting point to find additional 'mission design' related content, using the hyperlinks interconnecting the web pages.

From the initial 'Spacecraft design pages', six hyperlinks, judged as most relevant to a space mission corpus, were manually selected. These web pages were then automatically scrapped using the Python Selenium library, leading to the discovery of 1,023 additional non-redundant hyperlinks. The distribution of hyperlinks per web pages, including the main page on Spacecraft design, is shown in Figure 2. The list of web pages to be included in the corpus was manually filtered, for relevance to the project scope, and eventually yielding a corpus of 259 web pages. The content of each web pages was extracted and parsed with the Tika library [6], and saved as JSON files.

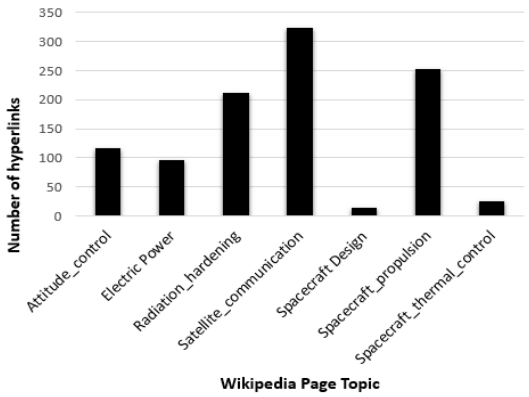

Figure 2: Wikipedia Corpus Distribution Pre-Filtering

3.2 *Application Corpus*

The application presented in this study focuses on the automatic categorisation of space mission requirements. The corpus of unseen documents, or requirements, submitted to the trained LDA model, is based on a set of 100 requirements extracted from two European Space Agency (ESA) documents, publicly available, the SMOS mission System Requirement Document [5] and MarcoPolo-R's Mission Requirement Document [4].

The requirements within these documents are organised per subsystems; for instance, all power-related requirements are found under the chapter 'Power requirements'. Therefore, with each requirement, a subsystem or topics to which the requirement belongs to can be extracted and used as ground truth for the categorisation evaluation. The distribution of requirements per topics is displayed in Figure 3. From this corpus, the 68 requirements related to the 7 topics of *Attitude and Orbit Control System/Guidance*, *Navigation and Control* (*AOCS/GNC*), *Communication*, *Environment*, *On−board Data Handling* (*OBDH*), *Power*, *Propulsion*, and *Thermal* are used to evaluate the general LDA model. The *Ground Segment*, *Launch*, *Mission Analysis*, *Payload* and *OBDH* topics are used to generate and evaluate focused LDA models.

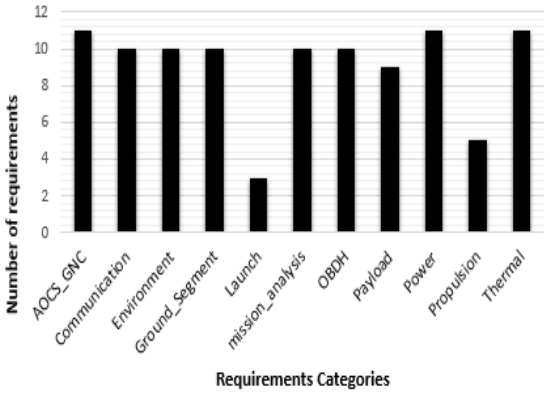

Figure 3: Application Corpus Distribution

4. METHODOLOGY

4.1 *Corpora Preprocessing*

A classic Natural Language Processing (NLP) language pipeline based on the Python NLTK (Natural Language Toolkit) library is used to preprocess both corpora. Each document, Wikipedia page or requirement, is tokenized, basic English stop words, as well as punctuation, numerical tokens, non-English characters, and urls are removed.A term frequency-inverse document frequency (tf-idf) analysis of the Wikipedia corpus is run to identify the tokens with the lowest score. The 10% of tokens with the lowest tf-idf are removed as tokens with low tf-idf have low informativeness value. To improve the topics dictionaries, the tokens corresponding to multiwords contained in the European Coordination for Space Standardization (ECSS) glossary of terms [3] are replaced within the corpora. In this study, the definition of multiwords is extended to concepts. To cover multiwords which are not found in [3], a manually validated list of multiwords related to space mission design is used. A collocation analysis over the Wikipedia corpus is also performed to find additional bigrams and trigrams. Table 1 displays a sample of multiwords found in the ECSS glossary and automatically found in the Wikipedia corpus via the collocation analysis. Finally, lemmatization is applied to the tokens to prevent grammatical redundancy within the topics dictionaries. In the case of the application corpus, acronyms found in requirements were expanded. Table 2 provides further information on the corpora statistics.

ECSS multiwords	End-of-life, Graveyard Orbit, Inertial Frame
Additional Bigrams	Earth Observation, Magnetic Field, Thermal Control
Additional Trigrams	Effective exhaust velocity, European Space Agency, Van Allen Belts

Table 1: Sample of multiwords identified within Corpus

Measure	**Wiki corpus**	**Requirement corpus**
Number of documents	259	100
Number of tokens	689,259	1,457
Size	9.6 MB	17 KB
Average tokens number	2,641	15
Dictionary size	36,031	577

Table 2: Corpora Statistics

4.2 *LDA Model Optimisation Process*

There are three main inputs to generate a model:

- the dictionary, which maps words to their identification numbers,
- the corpus, or document-term matrix, which provides per document, the words identification numbers and their frequency found within the document,
- the number of latent topics to be found

Other inputs, such as the number of passes, the number of times the LDA process goes through the entire corpus,

are arbitrarily set. The first two inputs are generated from the corpus. To choose the best number of topics, several LDA models with different numbers of topics are generated with the Gensim Python library [11]. The evaluation metric of perplexity, presented in the next paragraph, is used to determine which model is best fitted to represent the corpus topics distribution.
The Corpus is split between a training and a testing set, following the classic 80%/20% partition. The testing set will be used for the final evaluation of the retained model selected after the optimisation. The training set is used to perform a 5-fold cross-validation to find the number of the optimal topics and retain a potential final model. The optimisation process is summarised in Figure 4.

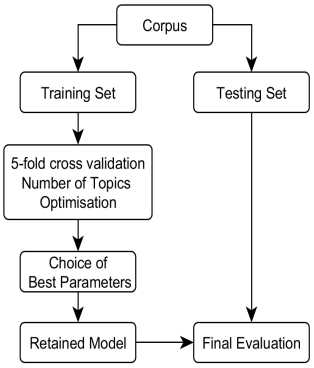

Figure 4: LDA Model Optimisation Process

4.3 *LDA Model Evaluation*

4.3.1 *Perplexity*

Perplexity is an intrinsic evaluation metrics used to evaluate LDA topics [2, 12]. Perplexity evaluates how well the probability distribution generated represents the corpus and measures the likelihood that the model will perform well with unseen, new, data. The value of perplexity must be minimised. Based on [2], perplexity over a test corpus, $per(D_{test})$ is expressed as:

$$per(D_{test}) = exp\left(-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right) \qquad (2)$$

with M, the number of documents in the test sample, w_d the words in document d, $\log p(w_d)$ the log likelihood of document d, and N_d, the number of tokens in document d.

4.3.2 *Latent Topics Visualisation*

LDAvis is a web-based interactive visualisation of LDA generated topics introduced in [13] to support the visualisation of LDA topics. The equivalent Python library is called pyldavis. This visualisation both provides an overview of the topics distribution relative to one another as well as deeper insights into the dictionaries of each topic. On the visualisation (as seen in Figure 6), the bigger the circle representing a topic, the more prevalent it is. A good LDA model should have big, non-overlapping topic circles scattered through space. The retained model should, therefore, have a minimised perplexity value as well as a satisfying topics distribution.

4.4 *Topics Labeling*

The topics dictionaries provided by the model are not labelled. Three human annotators have been involved in assigning topics labels to the dictionaries independently and manually. The annotators were given the following labels selection to choose from: *AOCS/GNC*, *Communication*, *Environment*, *Ground Segment*, *Launch*, *Mission Analysis*, *OBDH*, *Payload*, *Power*, *Propulsion*, and *Thermal*. The annotators also had the option to propose a topic label outside of this selection. It is made clear to the annotators that they can associate one to several labels to each dictionary and that one label can be associated with several dictionaries.

4.5 *Model Update for Specific Application*

A trained LDA model can be updated with an additional corpus of documents. Updating a trained model allows for fine-tuning it to more specific applications. The new corpus is transformed into the document-term matrix format with the same dictionary used to map the words to their id during the general model training. Updated LDA models mean updated topics and topics dictionaries. As in 4.4, the dictionaries are labelled by Human annotators for each update.

4.6 *Choice of seed words for semi-supervised LDA*

Seven sets of space mission design seed words are defined in an attempt to steer the model towards the following seven topics: *AOCS/GNC*, *Communication*, *Environment*, *OBDH*, *Power*, *Propulsion*, and *Thermal*. Each set is composed of around 20 words, each word can only belong to one set to avoid topics overlap. The seed words selected are based on a list of keywords or relevant concept associated to each topic. The list is validated by the same human annotators who performed the manual labeling. The selected seed words are presented in Table 3. To put the number of seed words given to the model into perspective, the topic dictionary being based on an initial set of 36,031 words, the seed words of each topic only represent 0.06 of the initial dictionary. To train the semi-supervised LDA model, the same number of topics as for the unsupervised model will be used. The unsupervised and semi-supervised model are both trained with the same Wikipedia-based corpus.

Topic Label	Space Seed words
AOCS/GNC	attitude, attitude control, guidance, navigation, reaction wheel, wheel, momentum, angular, body, freedom, gyroscope, motion, torque, torquer, star tracker, spin stabilised, stabilisation, sensor, gravity gradient, magnetotorquers
Communication	satellite communication, communication, band, bandwidth, packet, x band, transmitter, receiver, ka band, c band, frequency, antenna, relay, s band, l band, telemetry, tracking, telecommand, reception, command
Environment	environment, radiation, gamma ray, gamma radiation, particle, shield, dose, ray, shielding, electron, geomagnetic, van allen, single event upset, protection, cosmic, single event, space debris, debris, charging, background
OBDH	data handling, data rate, memory, storage, dram, sram, gbit, data, bitrate, cpu, ram, tag, encoder, decoder, downlink, uplink, computer, bit, measurement, execution, instruction, operation, processor
Power	power, battery, cell, solar cell, photovoltaic, solar power, voltage, watt, current, charge, discharge, power supply, battery powered, primary, secondary, lithium, circuit, energy, cycle, depth of discharge
Propulsion	propulsion, propulsion system, spacecraft propulsion, propellant mass, delta v, thruster, engine, propellant, ion, plasma, sail, electric, electric propulsion, nuclear, thrust, fuel, isp, total impulse, impulse, exhaust
Thermal	thermal control, thermal control system, heat pipe, heat, temperature, radiator, insulation, cooling, thermal, louver, heating, degree, thermodynamics, multi layer insulation, coating, overheating, mirror, heater, reflector, reflective

Table 3: Set of seeds words per topics for semi-supervised LDA

4.7 *Methodology for Automatic Categorisation*

The dictionary used by the retained general LDA model to map words to their ids is re-applied to the new corpus. The new corpus document-term matrix is generated based on this dictionary. The topic distribution defined by the LDA model can then be applied to the input requirement or query. The output is a list of latent topics along with the query's probability to belong to each topic. Only the top two topics will be retained for the categorisation evaluation. Using the chapters, the requirements were extracted from as ground truth allows us to evaluate the topics recommended by the LDA model.

4.8 *Automatic Categorisation Evaluation*

4.8.1 *Accuracy score*

The accuracy score only takes into consideration the top topic proposed by the LDA model. If this topic matches the requirement's ground truth, then the matching is considered a success. The accuracy score is divided by the number of queries, requirements, submitted to the model. Therefore, the best performance corresponds to an accuracy score of 1.

4.8.2 *Mean Reciprocal Ranking*

The MRR takes into consideration the top n answers of the model to a query. The score is inversely proportional to the correct answer, topic rank, as shown in Equation 3:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{3}$$

with Q the number of queries, $rank_i$, the rank of the ground truth. Only the top two topics proposed by the model will be taken into consideration.

5. RESULTS

5.1 *Retained General LDA Model*

For the optimisation process, a range of topics number from 4 to 100 was taken into account. The average perplexity observed for the LDA models depending on their topics number is shown on Figure 5 up to 30 topics. Eventually, a topic number of 22 was chosen for the retained model. The choice was based on the perplexity result of the final evaluation, 1.6e-05, computed with the held-out part of the Wikipedia corpus. The visualisation of topics distribution with pyldavis confirmed that topic numbers above 22 would lead to over-fitted models. The distribution of the topics for the retained model is displayed in Figure 6, where the number of the topics are eventually associated with labels presented in Table 4. The topics labeled as *AOCS/GNC* (12, 19, 21) and as *Environment* (2,4,10) are located in the same area of the visualisation, and will likely lead to less accurate distinction in-between the two topics during the categorisation. Similarly, the

OBDH topic is surrounded by the three topics related to *Power* (3, 5, 9). Finally topic 15, *Communication*, is completely embedded in the topic identified as *Ground Segment*. However, this is balanced by the well defined second Communication topic, 7.

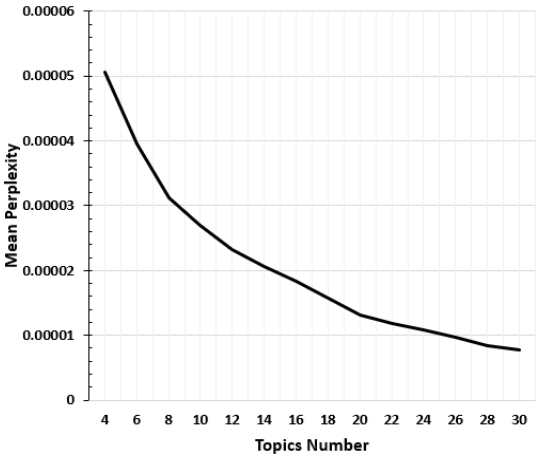

Figure 5: Average Perplexity for difference Topics Numbers

5.2 *Topics Dictionaries and Labeling*

Table 4 displays the result of the human annotation; only three topics could not be associated with any predefined labels. No case of several labels associated with one dictionary was reported. To label the topics, the annotators only took into consideration the most salient words: all words above the average of the top 50 words. Table 5 displays the dictionaries of each topics. The topics numbering is the same as the ones used in Figure 6.

Manual Label	Topic Number
AOCS/GNC	12, 19, 21
Communication	7, 15
Environment	2, 4, 10
Ground Segment	16
Launch	6, 11
Mission Analysis	8
OBDH	14
Payload	13
Power	3, 5, 9
Propulsion	20
Thermal	17

Table 4: Manual Labels and corresponding predominant topics (1 to 1). Topic 1, 18 and 22 are not assigned to any relevant topics.

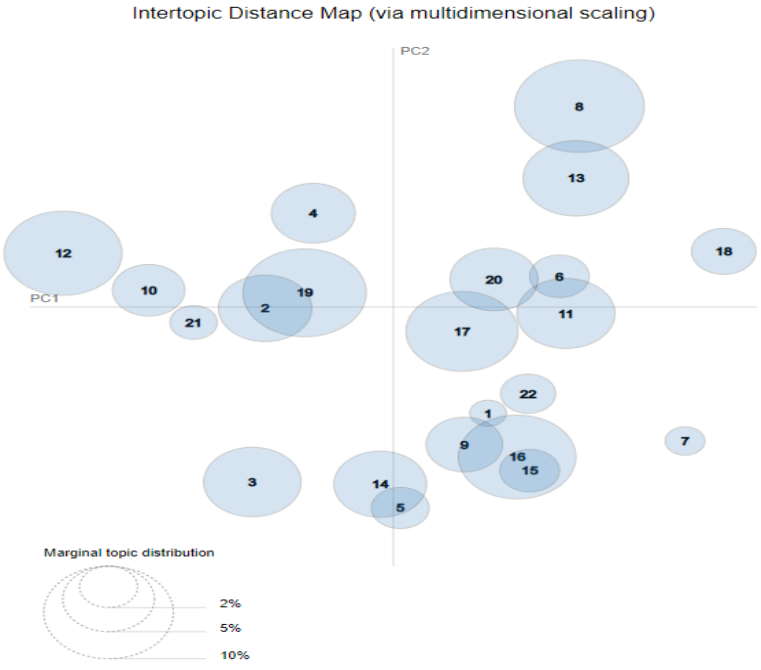

Figure 6: Latent Topics Visualisation for retained model

Manual Label	Topic Number	Dictionary
Social (other)	1	attitude, smart-1, processing, coulomb, cognitive, people, message, behavior, primestar, tv, model, dissonance, individual, belief, social, communication
Environment	2	radiation, gamma, ray, electron, transistor, nuclear, particle, decay, dose, ionization, base, photon
Power	3	capacitor, circuit, voltage, capacitance,dielectric, frac, board, resistance, layer, charge, plate, copper, frequency
Environment	4	cosmic, ray, universe, radiation, background, microwave, particle, temperature, physic
Power	5	battery, cell, rechargeable, discharge, voltage, charge, rate, lithium, capacity, storage
Launch	6	debris, orbital, leo, object, collision, kessler, junk, stage, mi, launch
Communication	7	anik, communication, canada, satcom, canadian,launch, launched, service, telesat, transponder,f1, television, hughes, network, series, cassiope
Mission Analysis	8	saturn, voyager, mar, probe, cassini, moon, jupiter, orbiter, lunar, pioneer, kosmos, galileo,titan, program, image, planet, flyby, instrument, launch
Power	9	cell, photovoltaic, efficiency, electricity, silicon, pv, plant, photovoltaics, cost, band, renewable, water, panel, grid, storage
Environment	10	magnetic, pole, magnet, line, model, magnetosphere, charge, geomagnetic, wind, dipole, north
Launch	11	gps, geostationary,receiver, launch, signal, service, united, global, communication, position, launched, orbital, aerospace, positioning, kosmos, military, error
AOCS/GNC	12	momentum, angular, velocity, body, motion, law, frac, mechanic, newton, vector, particle, object, rotation, combustion, equation
Payload	13	telescope, hubble, iridium, astronomy, launch, star, instrument, jwst, infrared, observatory, webb, mirror, galaxy, james, observation, object
OBDH	14	memory, dram, computer, cpu, logic, transistor, instruction, ram, cell, digital, operation, chip, bit, circuit, sram
Communication	15	antenna, network, packet, switching, communication, element, wave, transmission, impedance, internet
Ground Segment	16	radio, frequency, receiver, signal, communication, band, station, wave, antenna, network, transmitter, bandwidth
Thermal	17	heat, pipe, temperature, thermal, air, apollo, heating, tether, lunar, water, heater, gas, fluid, electron
Missions (Other)	18	esa, european, agency, launch, station, program, member, note, flight, shuttle, spaceflight
AOCS/GNC	19	sensor, tether, attitude, magnetic, orbital, magnetometer, gravity, quantum, wheel, laser, velocity, gyroscope, axis, reaction, momentum, bearing, specific
Propulsion	20	propulsion, thruster, engine, plasma, nuclear, sail, ion, thrust, propellant
AOCS/GNC	21	frame, inertial, relativity, coordinate, law, conservation, motion, equation, physic, acceleration
Distance (other)	22	cable, coaxial, astronomical, distance, intelsat, shield, line, kosmos, launch, conductor, length, light-year, communication

Table 5: Manual labelling of final model latent topics and corresponding dictionaries

5.3 *Automatic Categorisation with General LDA*

As previously mentioned in 3.2, 68 requirements related to the topics of *AOCS/GNC*, *Communication*, *Environment*, *OBDH*, *Power*, *Propulsion*, and *Thermal* were submitted to the retained LDA model, Figure 7 displays the accuracy score, and MRR obtained per requirement categories. The categorisation performance varies depending on the query topic, peaking at a score of 0.9 for the *AOCS/GNC* topic to lower scores of 0.1 for the *Communication* topic. Table 6 displays samples of successful and failed categorisation of requirements. The performances of the categorisation heavily depend on the initial corpus used to train the LDA and on the retained LDA model. The retained model used for this set of results could be considered as a general space mission design LDA as it is not focusing on any topics but is rather based on a corpus spreading over several topics.

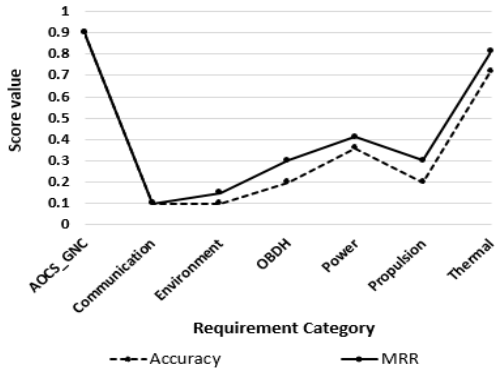

Figure 7: Automatic Categorisation with General LDA

Requirement	Ground Truth	LDA Model Output
The DHS system shall be compatible with the maximum data rates of each instrument as specified in RD16.	OBDH	(OBDH, 0.70), (missions, 0.17)
The TCS shall provide the appropriate thermal environment to the structural parts so that the alignment between sensors and instrument is maintained and the stability of the alignment is ensured.	Thermal	(thermal, 0.38), (AOCS GNC, 0.34)
Electrical power shall be guaranteed by a solar generator, its electrical configuration shall be defined on the basis of the topology selected for the EPS.	Power	(AOCS GNC, 0.42), (thermal, 0.24)
The performances of the propulsion system in terms of total impulse and margin shall satisfy the requirements imposed by the mission, the trajectory analysis and the overall system requirements.	Propulsion	(AOCS GNC, 0.61), (OBDH, 0.15)

Table 6: Successful and Failure Examples of Requirements Automatic Categorisation

5.4 *Categorisation with Focused LDA*

Specialised corpus on *Ground Segment*, *Launch*, *Mission Analysis*, *Payload* and *OBDH* are used to update the general LDA model into five specific, focused LDA models. The five specialised corpora used to update the model consists of a couple of Wikipedia web-pages, in average three web pages per category, parsed and pre-processed as the other corpora. Each updated LDA model generates new topic dictionaries and therefore requires new labelling. Human annotators are once again involved to label the new dictionaries. Table 7 illustrates the output of the labelling before and after the update. Each line corresponds to one focused LDA model, the focus being on a different category each time. Since the model is steered towards a specific category by the added corpus, more dictionaries tend to represent the category of interest.
Finally, the accuracy score and MRR are computed for each category on the general LDA model, and its corresponding focused LDA model. The results are respectively presented in 8 and 9. The categories used have low to average performances with the general LDA model. A significant increase in performance is observed with the specific LDA models. The sizes of each update corpus is a fraction of the training corpus size used to train the model. Therefore, the general LDA model presented in this study is an efficient base to develop focused LDA models for specific applications.

Topic Focus	**Topic Number in general LDA model**	**Topic Number in focused LDA model**
Ground Segment	16	3, 6, 10, 11, 12, 15, 16
Launch	6, 11	13, 17, 18, 19
Mission Analysis	8	8, 11, 12, 19
Payload	13	19
OBDH	14	11, 12, 14, 21

Table 7: Manual Labels and corresponding predominant topics before and after the general model update

Topic	**Average**	**Ground Segment**	**Launch**	**Mission Analysis**	**OBDH**	**Payload**
General LDA	0.23	0	0	0.4	0.2	0
Focused LDA	0.74	0.9	0.67	0.9	0.8	0.44

Table 8: Comparison of Accuracy Scores for a general and specific LDA models

Topic	**Average**	**Ground Segment**	**Launch**	**Mission Analysis**	**OBDH**	**Payload**
General LDA	0.12	0	0	0.4	0.3	0
Focused LDA	0.78	0.9	0.67	0.9	0.8	0.61

Table 9: Comparison of MRR for a general and specific LDA models

5.5 *Categorisation with semi-supervised LDA*

To evaluate the performance of the semi-supervised, or guided, LDA, 10 semi-supervised models are trained with the same corpus and the same number of topics as for the unsupervised retained model. Table 10 presents the percentage of seed words which were found into the final dictionaries top 20 words of one of these models. The percentages vary in-between the topics, but each topic seems to present one 'strong' dictionary more influenced than other dictionaries. Topics such as *Mission Analysis*, *Ground Segment* or *Launch* which had appeared in the general model are not prevalent anymore. The priors have therefore successfully focused the model on the 7 topics which were seeded.

Manual Label	**Topic Number**	**Percentage of priors found in dictionary top 20 words**
AOCS/GNC	1, 5, 12, 21	15, 20, 25, 5
Communications	8, 11, 14, 17	0, 20, 10, 30
Environment	3, 7, 9, 18	0, 15, 5, 40
OBDH	6, 10, 22	5, 0, 50
Power	4, 13, 15	15, 15, 5
Propulsion	2, 19	45, 20
Thermal	20	20

Table 10: Percentage of seed words found within topics dictionary

Figures 9 and 11 display a comparison of the accuracy score and of the MRR obtained with the general unsupervised and the semi-supervised models. For the semi-supervised models, the median of the results obtained with the 10 models was used. An improvement is noted for most of the topics, although the improvement rate varies. The semi-supervised models seem to have slightly homogenised the performances over all categories. To understand the performance distribution in-between the semi-supervised models, the box plots for the accuracy scores and MRR are respectively displayed in Figures 9 and 11.

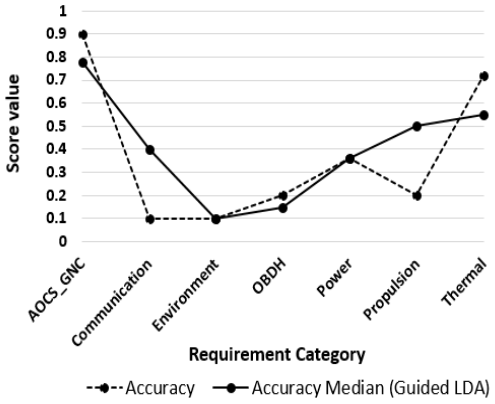

Figure 8: Comparison of categorisation accuracy score for unsupervised and semi-supervised LDA

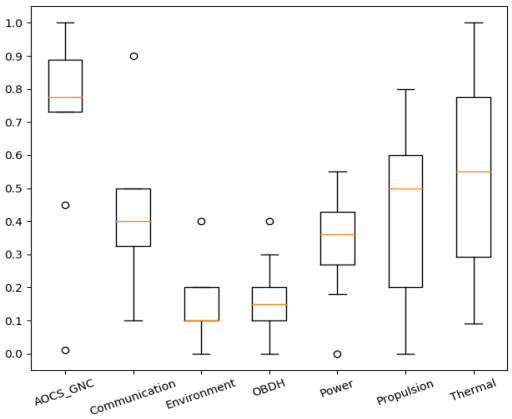

Figure 9: Accuracy score box plot for semi-supervised models

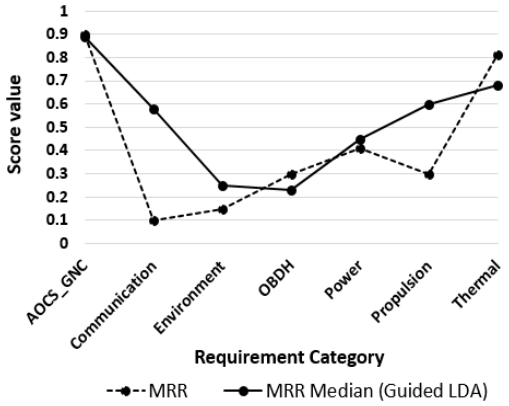

Figure 10: Comparison of categorisation MRR for unsupervised and semi-supervised LDA

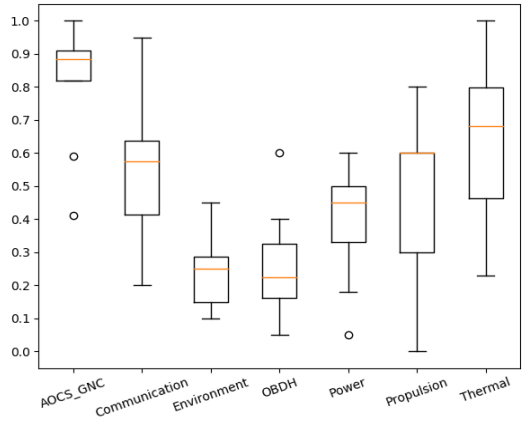

Figure 11: MRR box plot for semi-supervised models

6. Conclusion

This study presented the first application of LDA, a method of Topic Modeling, on a space mission design corpus. An extensive Wikipedia-based corpus focused on space mission design was collected, parsed, preprocessed, and used to train a general model. The LDA model was optimised based on the perplexity measure for a range of topics numbers. The topics dictionaries were labelled by human annotators. The result is a freely accessible general Space Mission Design LDA model. The performances of the model were evaluated with an automatic categorisation task. The categorisation performances of the general LDA model varied depending on the query categories. The performance was improved with a semi-supervised LDA model fed with lexical priors. The study also proved that the general LDA model could be adapted to specific categorisation tasks, pushing forward topics which were not previously salient and yielding high accuracy scores and MRR. The model and corpora are avail-

able at https://github.com/strath-ace/smart-nlp.
To further improve the general LDA model, an additional preprocessing step based on Part-Of-Speech could be implemented, to restrain the topics dictionaries to nouns. To train a model encompassing more space mission design topics, a wider and more complete corpus is needed. The lack of a complete freely-available 'Space Mission Design' corpus and the computational power required to train the LDA model are currently the main obstacles to the development of such a model.

Acknowledgement

This study was completed in the frame of the Design Engineering Assistant (DEA) project, a virtual assistant to support knowledge management and reuse, and decision-making at the early stages of space mission design [1]. The DEA is developed in the frame of an ESA Networking Partnership Initiative (NPI), the authors would like to warmly thank their partners for their valuable support: ESA, RHEA Systems, AIRBUS and satsearch.

References

[1] Audrey Berquand et al. "Artificial Intelligence for the Early Design Phases of Space Missions". In: *IEEE Aerospace*. Big Sky, Montana, US, 2019, pp. 1–20. ISBN: 9781538668542.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[3] ECSS. *ECSS Terms and Definition*. https://ecss.nl/home/ecss-glossary-terms/. 2007.

[4] ESA. *MarcoPolo-R Mission Requirements Document*. https://sci.esa.int/web/marcopolo-r/-/51297-marcopolo-r-mission-requirements-document/. 2012.

[5] ESA. *SMOS Systems Requirements Document*. https://earth.esa.int/web/guest/-/smos-system-requirements-document-6208. 2005.

[6] The Apache Software Foundation. *Apache Tika 1.20*. https://tika.apache.org/1.20/index.html.

[7] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. "Incorporating Lexical Priors into Topic Models". In: *13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France, 2012, pp. 204–213.

[8] Lucas Layman et al. "Topic Modeling of NASA Space System Problem Reports". In: *IEEE/ACM 13th Working Conference on Mining Software Repositories*. Austin, TX, US, 2016. DOI: http://dx.doi.org/10.1145/2901739.2901760.

[9] Yashar Moshfeghi, Benjamin Piwowarski, and Joemon M. Jose. "Handling Data Sparsity in Collaborative Filtering using Emotion and Semantic Based Features". In: *SIGIR'11*. Beijing, China, 2011.

[10] Ju Seop Park et al. "A new forecasting system using the latent dirichlet allocation (LDA) topic modeling technique". In: *WSEAS TRANSACTIONS on ENVIRONMENT and DEVELOPMENT* 14 (2018). ISSN: 2224-3496.

[11] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[12] Alexey P. Shiryaev et al. "LDA models for finding trends in technical knowledge domain". In: *2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017*. 2019. ISBN: 9781509048656. DOI: 10.1109/EIConRus.2017.7910614.

[13] Carson Sievert and Kenneth Shirley. "LDAvis: A method for visualizing and interpreting topics". In: *ACL Workshop on Interactive Language Learning, Visualization, and Interfaces*. https://github.com/bmabey/pyLDAvis. Baltimore, Maryland, US: Association for Computational Linguistics, 2014, pp. 63–70.