Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

STEAME ACADEMY
TEACHING FACILITATION LEARNING & CREATIVITY PLAN (L&C PLAN) - LEVEL 2
SERVICE TEACHERS: Text Mining: are these documents the same?

S T Eng A M Ent

● ● ●

1. Overview

Title	Text Mining: are these documents the same?
Driving Question or Topic	How can search engines find results for a user search based on keywords? How do computers identify text documents focusing on the same topics? How do computer algorithms model unstructured data for digital processing?
Ages, Grades, …	16 to 18 years old 10th to 12th grades
Duration, Timeline, Activities	18 hours 9 sessions of 2 hours each 21 activities
Curriculum Alignment	Data mining, machine learning, unstructured data modelling, computer programming
Contributors, Partners	
Abstract - Synopsis	Students are introduced to Data Mining and Machine Learning focusing on the core topics of digital processing of text. Text similarity is explored showing its mathematical foundations, intersection of sets and cosine between two vectors. Students work in teams to implement a simple tool to measure the similarity between two text documents. In the last sessions students are challenged for a competition to identify the best implementation. During all the sessions, students are being introduced to the key methods for text pre-processing, like stop-words and stemming. The last session is concluded by pushing the students to discuss and identify the resemblances between their implementation and a search engine and, from there, to design a search engine using their former implementation for text similarity.
References, Acknowledgements	

2. STEAME ACADEMY Framework*

Teachers' Cooperation	Teacher 1 (Science) • Machine learning and data mining: overview of the field, general architecture, common applications to daily life, common issues • Data modelling for digital processing: structured data representation for machine learning and data mining, unstructured data modelling for machine learning and data mining • Text mining: overview of the field, core concepts, modelling (Boolean model, TF model, TFxIDF model), similarity of text documents, pre-processing, main tasks and applications Teacher 2 (Engineering) • Python programming for text mining • Vector algebra in Excel Teacher 3 (Mathematics) • Cosine function, vector algebra • Set operators, intersection Teacher 1 cooperates with Teacher 2 and Teacher 3 to: - identify the Python libraries to use for text mining - identify the Excel functions to use for vector algebra - create the exercises and the challenge Teacher 1 cooperates with Teacher 2 to: - gather the needed corpora for the practical activities - annotate the needed corpora for each exercise Teacher 1 cooperates with Teacher 3 to: - introduce set operators in a text processing environment - introduce vector algebra in a text processing environment.
STEAME in Life (SiL) Organization	The last session is concluded by pushing the students to discuss and identify the resemblances between their implementation and a search engine, such as Google, and, from there, to design a search engine using their former implementation for text similarity.
Action Plan Formulation	**Preparatory Phase** 1. Research data mining and machine learning traditional and state-of-the-art

applications; link to search engines and generative AI; refer to cases of structured and unstructured data; review main challenges of data mining and machine learning (data modelling, multidimensionality, overfitting, missing data, data volume, big data, explaining versus predicting, …)
2. Gather and annotate corpora for exercises
3. Setup the Python programming environment (docker, repository in Github to clone, other)

Workshop Structure

1. Introduction
   1.1. Overview of Data Mining and Machine Learning: historical perspective, tasks/problems, applications, direct the talk to data modelling (start by tabular datasets and show examples, then ask about solutions for unstructured data such as images and text). Discuss these solutions with students.

2. Text modelling
   2.1. Explain the corpora we will use (must have small documents with a very reduced lexicon; include documents where TF can make a difference when compared to Boolean model).
   2.2. Refer again to text modelling and start with the Boolean model, show examples, ask how to find similarities between documents?
   2.3. Introduce set operators (union, intersection, etc.).
   2.4. Ask students in teams of 3 or 4 to implement a tool in Excel to implement their solutions and test them; each group explains which set operators they are using and why.
   2.5. Introduce vector algebra: internal product and cosine. Students discuss and try to find relations between set operators and vector algebra (intersection and internal product, union and sum). Refactor their Excel implementation to use now a combination of set and vector operators.
   2.6. Show the impact/relevance of term frequency in comparison to Boolean; students refactor to implement TF model and find similarities using the same corpora as before. It should now be clear that the cosine works fine.
   2.7. Discuss the changes and the problem that might arise when similarity is based on TF only (terms that are present in all documents have no discriminatory power); ask students to find solutions for this problem.
   2.8. Introduce IDF and TFxIDF; discuss, design and implement in Excel a similarity measure based on the TFxIDF model to be used for toy documents with a reduced lexicon.

3. Implementation
   3.1. Introduce R, Python or other libraries for text mining.
   3.2. Implement a function to compute the similarity between two documents using R, Python or other. All students' teams shall have a running function at the end of this phase; or else provide a basic implementation for all.

4. Mathematical Exploration
   4.1. Engage students in hands-on activities exploring set operators and vector algebra to compute the similarity between text documents in a corpus.
   4.2. Facilitate discussion on the mathematical principles behind text processing.

5. Culminating projects
   5.1. Students' teams are challenged to find, refactor and implement the best algorithm for a similarity function.
   5.2. Give students a corpus previously annotated for similarity between the static test queries/documents (each document in the corpus will be annotated with the similarity ranking for each one of the test cases), reserve a validation set (annotated as well; this can be two or three text documents, each one with a few terms as if it was a query for a search engine; for each one of these "queries" provide the ranking of the documents in the students corpus, this will be used to compute F1 in the end and announce the winner).
   5.3. Provide time for students to implement and fine tune their similarity function to get the best results on their corpus, incorporating feedback and guidance from facilitators who will introduce students to pre-processing techniques while the discussion progresses (stop-word removal, stemming, etc.).

6. Link to search engines
   6.1. A "document" can be a query as the ones we use in search engines like Google, such as "camélias porto" or "computer vision". Show in real using Google. Give students the validation set, i.e., three documents, each one with a few keywords, as if they were a query for a search engine; run the similarity functions to provide a ranking of the documents in the corpus and give students the best ranking. Compute the winner using F1 measure. Explain students what is F1, Recall and Precision.

Evaluation and Reflection

1. Assess students' understanding and application of set and vector algebra operators and concepts through project-based assessments, presentations, and written reflections.
2. Encourage students to reflect on their learning experiences, highlighting the relationship between mathematics and text mining.

Ask students to design and present a search engine methodology and a non-functional prototype using what they have learned.

[*] *under development the final elements of the framework*

3. Objectives and Methodologies

Learning Goals and	1. Understand the generic modelling and processing concepts and techniques

Objectives	used in text mining 2. Explore the interdisciplinary connections between text mining, search engines and vector algebra 3. Illustrate the similarity between text documents and other unstructured data sets as applications of vector algebra
Learning Outcomes and expected Results	**Learning Outcomes** A. Discuss high level topics related to the fields of text mining and search engines B. Describe the relation between vector algebra, set theory and text mining C. Apply basic techniques of text mining to address simple use cases **Expected Results** 1. Text document similarity function in R, Python or other programming language
Prior Knowledge and Prerequisites	1. Fundamental knowledge of vector algebra 2. Familiarity with Excel 3. Basic software programming skills 4. Proficient use of IT tools
Motivation, Methodology, Strategies, Scaffolds	1. Assign students to small teams (3 or 4 students). 2. Design a solution, implement, test and refine in an iterative way. Use an iterative development methodology. 3. Highlight the connections between vector algebra, document models, cosine and similarity. 4. Explore search engines to showcase the relations between text similarity and search engines' results. 5. Guide students through a evolutionary path from the most simple models (Boolean) to more complex approaches (TFxIDF), introducing challenges step by step while essay8ng in Excel with basic implementations for small documents with one or two short sentences and a few distinct terms, from a lexicon of 10 or 20 terms.

4. Preparation and Means

Preparation, Space Setting, *Troubleshooting Tips*	The workshop will be held in a classroom for approximately 20 students, in groups of 3 to 4 students. Ideally the classroom layout will be organized in 5 to 7 groups of tables where students from each team can sit facing each other. The room needs a beamer and a wall for presentations to all and a white board with pens to discuss ideas.
Resources, Tools,	A repository in GDrive, Teams, Github or any other provider shall be prepared in

Material, Attachments, Equipment	advance with all the programming environment (R, Python, …) and the corpora needed for the hands-on sessions. A document must be provided to guide students throughout all the course/workshop, explaining details, expected results, assessment and learning outcomes per session.
Health and Safety	

5. Implementation

Instructional Activities, Procedures, Reflections	**Workshop Structure** 1. Introduction [Session 1: 2 hours, 3 activities] 1.1. Overview of Data Mining and Machine Learning: historical perspective, tasks/problems, methodologies, techniques and tools (Weka, R, …), tasks (classification, clustering, …), applications, direct the talk to data modelling (start by tabular datasets and show examples). [40 minutes debate] 1.2. Debate with students about modelling of unstructured data (such as images and text) for automatic processing. Discuss alternative solutions with students while introducing relevant matters (bag of words, synonyms, location of text: title, abstract, …). [40 minutes, debate] 1.3. Introduce document similarity and its relevance in text mining. [40 minutes, debate] 2. Text modelling 1 [Session 2: 2 hours, 4 activities] 2.1. Refer to text modelling and start demonstrating a text similarity function with the Boolean model, show examples, ask how to find similarities between documents? [40 minutes demonstration, hands-on] 2.2. Explain the corpora we will use (must have small documents with a very reduced lexicon; include documents where TF can make a difference when compared to Boolean model). [20 minutes, expository session] 2.3. Introduce set operators (union, intersection, etc.). [20 minutes, expository, demonstration] 2.4. Ask students in teams of 3 or 4 to implement a text similarity function in Excel and test it; each group explains which set operators they are using and why. [40 minutes, hands-on] 3. Text modelling 2 [Session 3: 2 hours, 4 activities] 3.1. Introduce vector algebra: internal product and cosine. Facilitate discussion on the mathematical principles behind text processing. [20 minutes, expository, demonstration] 3.2. Engage students in hands-on activities exploring set operators and vector algebra to compute the similarity between text documents in a corpus. Students debate in teams to discuss/find relations between set operators and vector algebra (intersection and internal product, union and sum). Refactor Excel implementations of the similarity function to use a combination of set and vector operators. [40 minutes, expository session]

3.3. Show the impact/relevance of term frequency in comparison to Boolean when computing similarity between documents. [20 minutes, expository session]
3.4. Students refactor their implementation to implement TF model and find similarities using the same corpora as before. It should now be clear that the cosine works fine. [40 minutes, expository session]

4. Text modelling 3 [Session 4: 2 hours, 3 activities]
4.1. Discuss the changes and the problem that might arise when similarity is based on TF only (terms that are present in all documents have no discriminatory power). [20 minutes, expository session]
4.2. Students find solutions for this problem. [60 minutes, expository session]
4.3. Introduce IDF and TFxIDF. Discuss, design and implement with students a similarity measure based on the TFxIDF model to be used for toy documents with a reduced lexicon. [40 minutes, demonstration]

5. Implementation [Session 5: 2 hours, 3 activities]
5.1. Introduce R, Python or other libraries for text mining. [30 minutes, expository session]
5.2. Setup the development environment for text mining. Teacher have prepared a guide and provided it to students. [30 minutes, project work]
5.3. Students, in teams, implement a function to compute the similarity between two documents using R, Python or other. All students' teams shall have a running function at the end of this phase (teachers prepare in advance an implementation to provide for all if needed). [60 minutes, hands-on]

6. Culminating project 1 [Session 6: 2 hours, 1 activity]
6.1. Students' teams are challenged to find, refactor and implement the best algorithm for a similarity function. [120 minutes, research, project work]
7. Culminating project 2 [Sessions 7 and 8: 4 hours, 1 activity]
7.1. Give students a corpus previously annotated for similarity between the static test queries/documents (each document in the corpus will be annotated with the similarity ranking for each one of the test cases), reserve a validation set (annotated as well; this can be two or three text documents, each one with a few terms as if it was a query for a search engine; for each one of these "queries" provide the ranking of the documents in the students corpus, this will be used to compute F1 in the end and announce the winner). Provide time for students to implement and fine tune their similarity function to get the best results on their corpus, incorporating feedback and guidance from facilitators who will introduce students to pre-processing techniques while the discussion progresses (stop-word removal, stemming, etc.). [240 minutes, project work, hands-on]

8. Link to search engines [Session 9: 2 hours, 2 activities]

	8.1. A “document” can be a query as the ones we use in search engines like Google, such as “camélias porto” or “computer vision”. Show in real using Google. Give students the validation set, i.e., three documents, each one with a few keywords, as if they were a query for a search engine; run the similarity functions to provide a ranking of the documents in the corpus and give students the best ranking. Compute the winner using F1 measure. Explain students what is F1, Recall and Precision. [60 minutes expository session] 8.2. Debate, reflexion and conclusions. [60 minutes expository session]
Assessment - Evaluation	**Evaluation and Reflection** 1. Assess students' understanding and application of set and vector algebra operators and concepts through project-based assessments, presentations, and written reflections. 2. Encourage students to reflect on their learning experiences, highlighting the relationship between mathematics and text mining. 3. Ask students to design and present a search engine methodology and a non-functional prototype using what they have learned.
Presentation - Reporting - Sharing	1. A function that computes the similarity between two text documents. 2. A presentation in PowerPoint describing a methodology and a non-functional prototype of a novel search engine proposed by the students using what we have developed in the workshop (the document similarity function).
Extensions - Other Information	

STEAME ACADEMY Prototype/Guide for Learning & Creativity Approach
Action Plan Formulation

Major steps in the STEAME learning approach:

STAGE I: Preparation by one or more teachers

1. Formulating initial thoughts on the thematic sectors/areas to be covered
2. Engaging the world of the wider environment / work / business / parents / society / environment/ ethics
3. Target Age Group of Students - Associating with the Official Curriculum - Setting Goals and Objectives
4. Organization of the tasks of the parties involved - Designation of Coordinator - Workplaces etc.

STAGE II: Action Plan Formulation (Steps 1-18)

Preparation (by teachers)

1. Relation to the Real World – Reflection
2. Incentive – Motivation
3. Formulation of a problem (possibly in stages or phases) resulting from the above

Development (by students) – Guidance & Evaluation (in 9-11, by teachers)

4. Background Creation - Search / Gather Information
5. Simplify the issue - Configure the problem with a limited number of requirements
6. Case Making - Designing - identifying materials for building / development / creation
7. Construction - Workflow - Implementation of projects
8. Observation-Experimentation - Initial Conclusions
9. Documentation - Searching Thematic Areas (AI fields) related to the subject under study – Explanation based on Existing Theories and / or Empirical Results
10. Gathering of results / information based on points 7, 8, 9
11. First group presentation by students

Configuration & Results (by students) – Guidance & Evaluation (by teachers)

12. Configure STEAME models to describe / represent / illustrate the results
13. Studying the results in 9 and drawing conclusions, using 12
14. Applications in Everyday Life - Suggestions for Developing 9 (Entrepreneurship - SIL Days)

Review (by teachers)

15. Review the problem and review it under more demanding conditions

Project Completion (by students) – Guidance & Evaluation (by teachers)

16. Repeat steps 5 through 11 with additional or new requirements as formulated in 15
17. Investigation - Case Studies - Expansion - New Theories - Testing New Conclusions

18. Presentation of Conclusions - Communication Tactics.

STAGE III: STEAME ACADEMY Actions and Cooperation in Creative Projects for school students

Title of Project: _____
Brief Description/Outline of Organizational Arrangements / Responsibilities for Action

STAGE	**Activities/Steps** Teacher 1(T1) Cooperation with T2 and student guidance	**Activities /Steps** **By Students** Age Group: ____	**Activities /Steps** Teacher 2 (T2) Cooperation with T1 and student guidance
A	Preparation of steps 1,2,3		Cooperation in step 3
B	Guidance in step 9	4,5,6,7,8,9,10	Support guidance in step 9
C	Creative Evaluation	11	Creative Evaluation
D	Guidance	12	Guidance
E	Guidance	13 (9+12)	Guidance
F	Organization (SIL) STEAME in Life	14 Meeting with Business representatives	Organization (SIL) STEAME in Life
G	Preparation of step 15		Cooperation in step 15
H	Guidance	16 (repetition 5-11)	Support Guidance
I	Guidance	17	Support Guidance
K	Creative Evaluation	18	Creative Evaluation