

Rapport Intermédiaire

PDF2txt

La première étape de traitement a été de convertir des fichiers pdf en format textuel. Afin d'automatiser la conversion, nous avons utilisé la librairie **pdfplumber**.

Un format .pdf n'est pas fait pour être converti en .txt, ainsi différents problèmes sont apparus dans la conversion :

- Certains éléments de mise en page sont considéré comme du texte et apparaissent dans les différents import :
- 32 Armenia Educational national plan.doc 2 State program for education

33 CONTENTS

34

35

36

37 Part 1. Situation Overview and Problems in the Field of Education

38

39 I. Brief Overview of the Education System.....4

40

41 II. Problems in the Field of Education.....11

42

43

44 Part 2. Program Objectives and Implementation Timelines20

45

46

47 Part 3. Program Activities.....22

48

49

50 Part 4. State and Social Guarantees for the Students31

51

52

53 Part 5. Program Funding.....32

54

55

56 Part 6. Program Activities and Implementation Timelines33

57

58

59 Attachments.....48

60 Armenia Educational national plan.doc 3 State program for education

61

CONTENTS

Part 1. Situation Overview and Problems in the Field of Education4

I. Brief Overview of the Education System.....4

II. Problems in the Field of Education.....11

Part 2. Program Objectives and Implementation Timelines20

Part 3. Program Activities.....22

Part 4. State and Social Guarantees for the Students31

Part 5. Program Funding.....32

Part 6. Program Activities and Implementation Timelines33

Attachments.....48
- Certains élément ne sont pas importés correctement, en particulier les tableaux et les sommaires, leur import comporte souvent des lignes morceaux de texte isolés.

- 393 • Unions and other

394 for appeals monitoring formal quality units' appraisal

395 professional

396 adopted by MOES guidelines

397 • Develop multiple level associations • Fear of retribution

398 approved and

399 of reporting on staff • Rigor and quality of for reporting low

400 • Regional MOES adopted by

401 performance self evaluation report quality and

402 and LGA staff 2006.

403 of departments performance levels

404 e • Develop evidence

405 nc • School Directors • 50% of

406 a based appraisal system • External audit by

407 ur and teachers supervisors and

408 s the MOES

409 As • Bench mark MOES HOD's adopt

410 uality pmeinfiosrtmiesan acned w ith other • General public

411 Q

412 d international best • 100% of

413 n

414 a practices.

415 g supervisors and

416 n

417 5 orti HOD's adopt

418 B.1. Rep tQhAe nbeyw 2 0SIA0 and

419

420 B.2 IMPROVING THE QUALITY OF THE TEACHING AND

421 LEARNING PROCESS: POLICY MATRIX

422 - 36 -Key Objectives Beneficiaries Monitoring Risks & Prop

423 issue indicators assumptions timeline

B.1.5 Reporting and Quality Assurance	<ul style="list-style-type: none">Develop staff appraisal systems with provision for appealsDevelop multiple level of reporting on staff performanceDevelop evidence based appraisal systemBench mark MOES performance with other ministries and international best practices.	<ul style="list-style-type: none">MOES staffUnions and other professional associationsRegional MOES and LGA staffSchool Directors and teachersGeneral public	<ul style="list-style-type: none">Structure and guidelines for monitoring formal adopted by MOESRigor and quality of self evaluation report of departmentsExternal audit by the MOES	<ul style="list-style-type: none">Agreeing on objective indicators of qualityFear of retribution for reporting low quality and performance levels	<ul style="list-style-type: none">Staff and organizational units' appraisal guidelines approved and adopted by 2006.50% of supervisors and HOD's adopt the new SA and QA by 2008100% of supervisors and HOD's adopt the new SA and QA by 2010
--	---	--	--	--	---

B.2 IMPROVING THE QUALITY OF THE TEACHING AND LEARNING PROCESS: POLICY MATRIX
- Enfin les termes numériques sont nombreux et difficilement analysables, les numéros de pages et autres annotations chiffrées sont plutôt inutiles pour notre problème.

La solution de facilité consistant à retirer les parties problématiques a été choisie puisque l'importation des données ne fait pas parti du PSC. Il aurait été souhaitable qu'un set d'entrainement soit fourni pour l'intégralité du sujet mais nous y reviendront ultérieurement.

Quelques requetes **REGEX** pour formater le texte fourni sont donc utilisés pour s'assurer que le traitement est bien effectué sur de texte mais des abérations subsistent néanmoins.

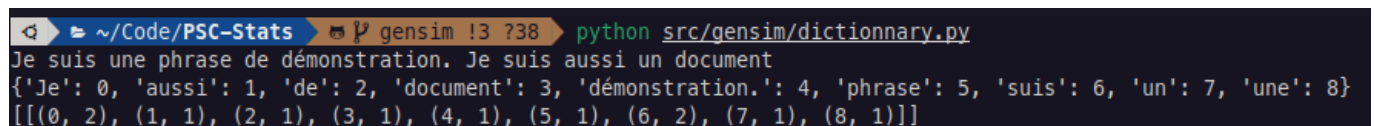
Certains mots sont importés en étant concaténés avec d'autres ce qui rend impossible leur utilisation et au contraire, certaines lettres sont isolés du reste des leurs rendant de nouveau impossible l'analyse. Nous avons de nouveau choisi d'éliminer les entrées non cohérents pour faciliter le traitement.

Cependant quelques difficultés subsistent : Il est difficile de déterminer automatiquement les mots à retirer (pour les lettres isolés ou les mots dépassant les 40 lettres certes mais le problème est plus complexe). Nous pensons ainsi à utiliser des dictionnaires des différents langages appréhendés (en l'occurrence anglais) et vérifier que les termes sont censés.

Doc2Vec

Bag Of Words (BOW)

Une fois les importations effectués, un premier modèle abordable est le **Doc2Vec**. On modélise chaque document du corpus par un simple vecteur qui formé de tous les mots qui le compose :



```
~/Code/PSC-Stats  P gensim 13 738  python src/gensim/dictionary.py
Je suis une phrase de démonstration. Je suis aussi un document
{'Je': 0, 'aussi': 1, 'de': 2, 'document': 3, 'démonstration.': 4, 'phrase': 5, 'suis': 6, 'un': 7, 'une': 8}
[[ (0, 2), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 2), (7, 1), (8, 1) ]]
```

Chaque mot du document est représenté par un couple : (ID, Nombre d'occurrence)

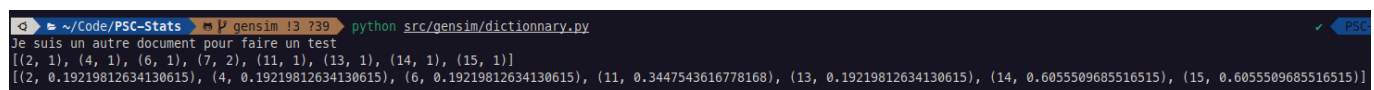
On construit donc de grands vecteurs pour représenter nos documents sans prendre en compte la position des mots dans le document et on peut faire de l'analyse statistique dessus.

Une des premières méthodes proposée est d'implémenter une métrique pour donner de l'importance aux différents mots : le **TF-IDF**

TF-IDF

Une première approche consiste à supposer qu'un terme présent de nombreuses fois définit correctement le texte doit avoir de l'importance mais il s'avère que les langues sont construites avec de nombreux mots de liaisons et très peu utiles pour définir le sens d'une phrase. On commence donc par retirer tous ces mots appelés **stopwords** pour faciliter l'analyse.

Cependant pour un corpus donné, il est intéressant de pénaliser des mots qui sont propres au contexte donné et non à un document particulier. On utilise donc la métrique **TF-IDF** pour *Term Frequency, Inverse Term Frequency*. On compte les occurrences d'un mot dans un document et également le nombre de documents où il apparaît pour lui donner un poids définitif.



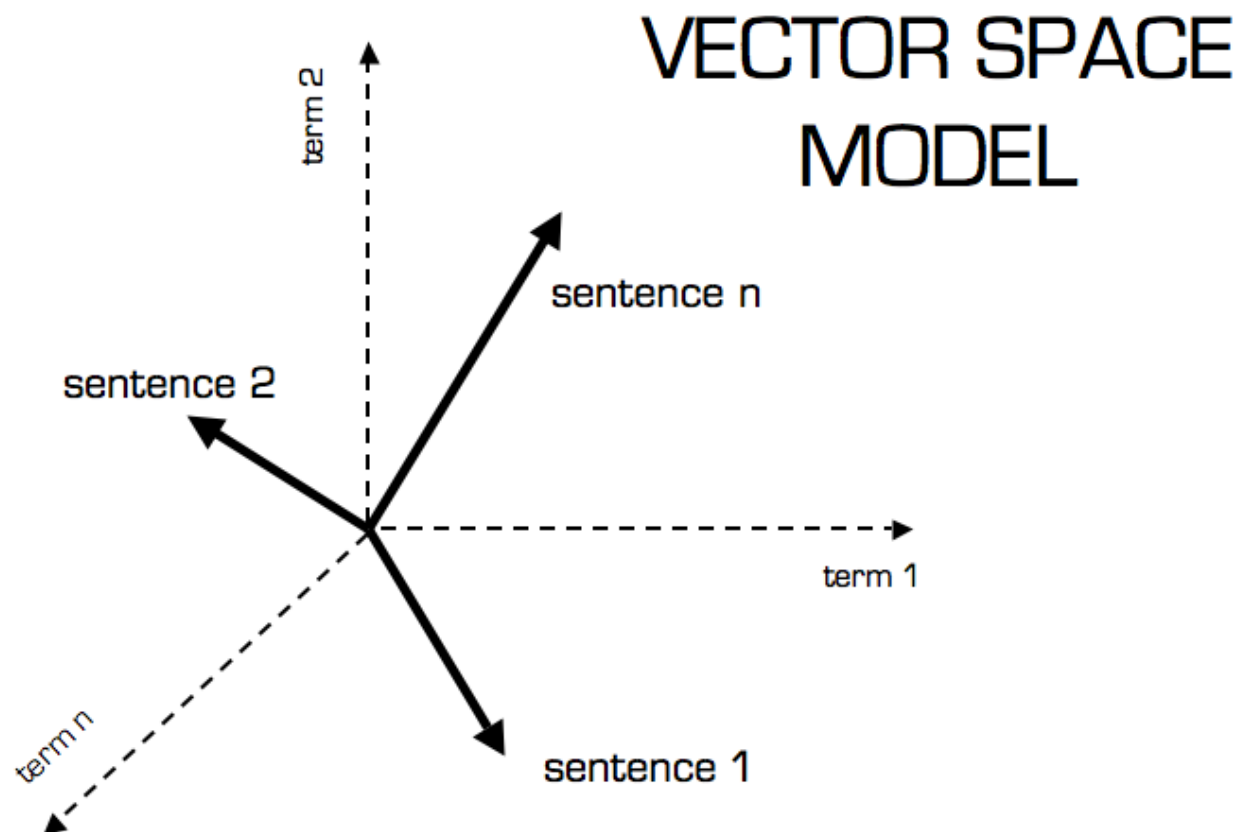
```
~/Code/PSC-Stats  P gensim 13 739  python src/gensim/dictionary.py
Je suis un autre document pour faire un test
[(2, 1), (4, 1), (6, 1), (7, 2), (11, 1), (13, 1), (14, 1), (15, 1)]
[(2, 0.19219812634130615), (4, 0.19219812634130615), (6, 0.19219812634130615), (11, 0.3447543616778168), (13, 0.19219812634130615), (14, 0.6055509685516515), (15, 0.6055509685516515)]
```

Ainsi on obtient des couples (ID, poids) pour quantifier l'importance des termes dans un texte

SIM

Enfin, avec cette métrique implémentée, il est possible de définir à quel points des textes se ressemblent (un % de ressemblance).

On utilise une **cosine-sim** : On réalise le produit scalaire de 2 vecteurs formant chacun un document et on compare les angles pour déterminer la ressemblance.



Cependant, chaque mot est défini de manière complètement orthogonale à ses pairs. La proximité entre des termes comme "homme" et "corps" est la même que "homme" et "éléphant" ce qui ne rend pas une description cohérente avec le sens des mots.

Pour remédier à cela, nous creusons actuellement les pistes du **Word2Vec** et **Latent Semantic Analysis** (LSA ou LSI) pour avoir une représentation plus fidèle des documents et des visualisation de leur proximité sur des graphes et non de simples "mesures d'angle".

Résultats

Le TF-IDF donne les termes les plus intéressant sur les différents corpus :

```

~/Code/PSC-Stats  gensim !3 ?42  python src/gensim/tfidf.py
armenia 0.39409267092970296
elaboration 0.39409267092970296
field 0.37720298503271565
roa 0.27023497435179633
doc 0.2589751837538048
mid 0.23645560255782178
quarter 0.20267623076384722
scientific 0.15763706837188118
upbringing 0.1294875918769024
pedagogues 0.12385769657790664
budgetary 0.11822780127891089
formation 0.10696801068091938
courses 0.10133811538192361
fields 0.09570822008292786
allocation 0.07881853418594059
cadres 0.07881853418594059
diaspora 0.07881853418594059
decreased 0.07318863888694484
needing 0.07318863888694484
attestation 0.06755874358794908

develop 0.5006062649583032
moes 0.40258545783360045
mths 0.2905616782625116
process 0.20304310047259846
central 0.16453492624503668
weeks 0.14353046757545754
monitoring 0.14002972446386103
wkshops 0.14002972446386103
cost 0.1260267520174749
alternative 0.11552452268268534
lg 0.11202377957108882
emis 0.10502229334789576
roles 0.09802080712470271
train 0.09802080712470271
resource 0.09452006401310618
model 0.08751857778991314
costs 0.08051709156672009
review 0.08051709156672009
use 0.08051709156672009
content 0.07351560534352704

```

Légende : Top 20 (TF-IDF) mots pour un corpus de 2 documents sur l'éducation -

On remarque cependant que les résultats sont peu lisibles puisque ce sont les racines qui sont analysés. Il est parfois difficile de trouver le sujet qui correspond au mot (mid pour mid-level professional education dans le premier ou MOES = Ministry of Education and Science dans le second)

Pour la similarité, on obtient les résultats suivants pour un corpus d'environ 20 documents :

```

Doc : montenegro_strategy_for_the_development_of_higher_education_2016-2020.txt
hungary_public_ed_dev_strategy_2004_en.txt 7%
estonia-higher-education-strategy-2006-2015.txt 7%
czech_republic_higher_education_strategic_plan_2021.txt 7%
netherlands_quality_in_diversity_strategy.txt 7%
denmark_better_education_action_plan.txt 6%

Doc : strategic_framework_cz20301.txt
czech_republic_higher_education_strategic_plan_2021.txt 35%
czech_republic_framework_education_programme_for_basic_education.txt 7%
hungary_public_ed_dev_strategy_2004_en.txt 7%
lithuania_ed_improvement_project_2002-2005.txt 7%
moldova_consolidated_strategy_ed-dev_2011-2015.txt 6%

Doc : denmark_better_education_action_plan.txt
czech_republic_higher_education_strategic_plan_2021.txt 13%
netherlands_quality_in_diversity_strategy.txt 12%
hungary_public_ed_dev_strategy_2004_en.txt 12%
estonia-higher-education-strategy-2006-2015.txt 6%
moldova_consolidated_strategy_ed-dev_2011-2015.txt 6%

Doc : ireland_department-of-education-and-skills-strategy-statement-2016-2019.txt
ireland_national_skills_stratgey_2025.txt 32%
ireland_statement-of-strategy-2019-2021.txt 23%
reportsstrategy-statementdepartment-of-education-and-skills-statement-of-strategy-2015-2017.txt 20%
ireland_strategy_statement.txt 15%
ireland_national-strategy-for-higher-education-2030-implementation-plan.txt 11%

Doc : georgia_consolidated_education_strategy_and_action_plan_2007-2011.txt
albania-education-strategy-2004-2015.txt 19%
moldova_education_plan2006-2008.txt 12%
croatia_education_sector_development_plan_2005-2010.txt 8%
moldova_consolidated_strategy_ed-dev_2011-2015.txt 8%
armenia_educational_national_plan.txt 8%

Doc : albania-education-strategy-2004-2015.txt
georgia_consolidated_education_strategy_and_action_plan_2007-2011.txt 19%
hungary_public_ed_dev_strategy_2004_en.txt 6%
moldova_consolidated_strategy_ed-dev_2011-2015.txt 5%
croatia_education_sector_development_plan_2005-2010.txt 5%
ireland_strategy_statement.txt 4%

```

On remarque une bonne cohérence des résultats pour certains domaines (les documents parlant du même pays sont bien regroupés ensemble) mais il est difficile d'apprécier la mesure de similarité puisqu'aucun travail d'analyse n'a été réalisé par un humain.

Ainsi, nous ne pouvons pas mesurer facilement la pertinence de nos modèles puisque les données de test ne sont pas assez adaptées à nos besoins.

Une décision commune a été prise avec le tuteur : nous effectuons nos recherches avec des datasets pensés pour ce genre de problèmes afin de perfectionner nos connaissances comme nos modèles et pouvoir interpréter nos résultats plus sereinement.

Nous revenons cependant de temps en temps aux corpus d'application proposés lorsque nous avons mieux appréhendés les différentes problématiques.

Annexes :

Enfin, une webapp est en cours de développement pour réaliser une démonstration des différentes techniques en temps réel. Voilà l'état actuel (réalisé avec la librairie **gradio**)

Demo

CorpusPDFTF-IDF

TF-IDF

Pick Corpus

DocA

Pick DocA

docA.pdf

Stats

Similarity %

56

DocB

Pick DocB

docB.pdf

Top Words DocA

Top Words DocA

- educ 585

- school 194

- institut 167

- state 159

- program 117

- the 111

- develop 109

- profession 98

- system 95

- secondari 86

Top Words DocB

Top Words DocB

- develop 303

- school 228

- educ 224

- teacher 144

- the 138

- moe 117

- train 115

- level 111

- curriculum 87

- system 86

Use via API

Built with Gradio

6 / 6