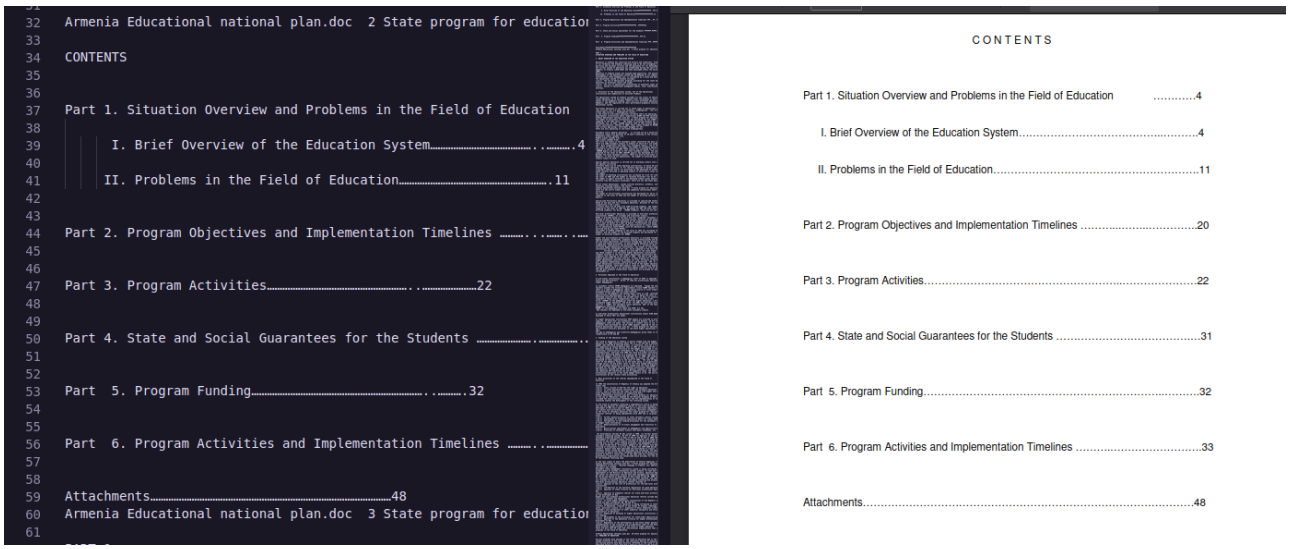
Rapport Intermédiaire

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 :

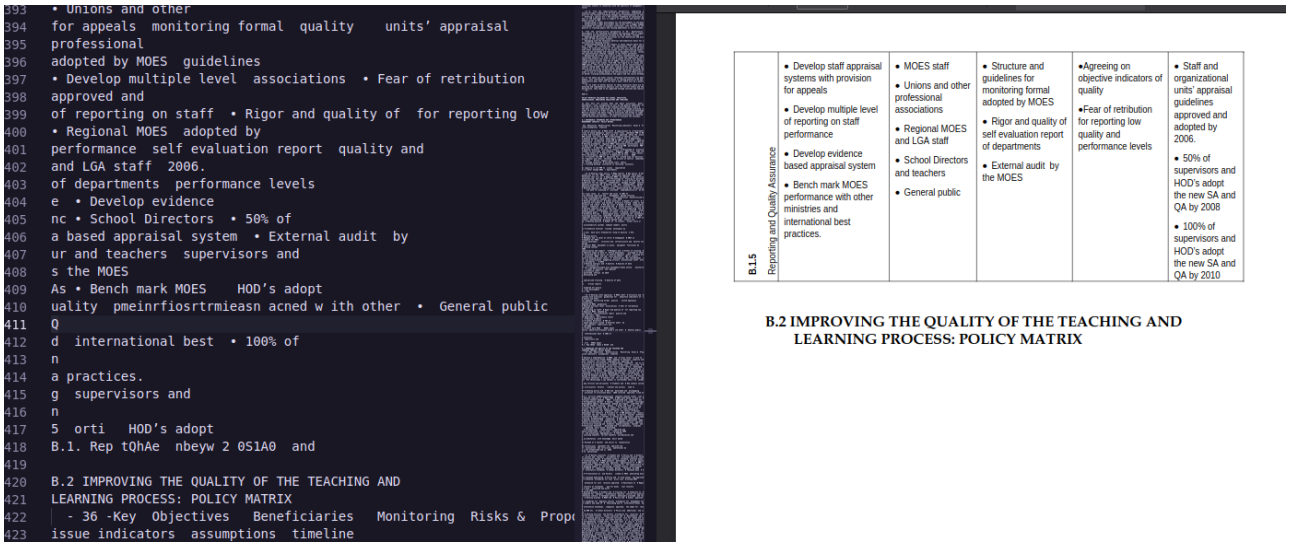
- Certains éléments de mise en page sont considéré comme du texte et apparaissent dans les différents import :



```
Armenia Educational national plan.doc  2 State program for educatio

CONTENTS

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

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

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

Part 2. Program Objectives and Implementation Timelines ……

Part 3. Program Activities……..22

Part 4. State and Social Guarantees for the Students ……

Part  5. Program Funding……..32

Part  6. Program Activities and Implementation Timelines ……

Attachments……48
Armenia Educational national plan.doc  3 State program for educatio
```

CONTENTS

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



```
• Unions and other
for appeals  monitoring formal  quality    units' appraisal
professional
adopted by MOES  guidelines
• Develop multiple level  associations  • Fear of retribution
approved and
of reporting on staff  • Rigor and quality of  for reporting low
• Regional MOES  adopted by
performance  self evaluation report  quality and
and LGA staff  2006.
of departments  performance levels
e  • Develop evidence
nc • School Directors  • 50% of
a based appraisal system  • External audit  by
ur and teachers  supervisors and
s the MOES
As • Bench mark MOES    HOD's adopt
uality  pmeinrfiosrtrmieasn acned w ith other  •  General public
Q
d  international best  • 100% of
n
a practices.
g  supervisors and
n
5  orti   HOD's adopt
B.1. Rep tQhAe  nbeyw 2 0S1A0  and

B.2 IMPROVING THE QUALITY OF THE TEACHING AND
LEARNING PROCESS: POLICY MATRIX
  - 36 -Key  Objectives   Beneficiaries   Monitoring  Risks &  Propo
issue indicators  assumptions  timeline
```

B.1.5 Reporting and Quality Assurance	• Develop staff appraisal systems with provision for appeals • Develop multiple level of reporting on staff performance • Develop evidence based appraisal system • Bench mark MOES performance with other ministries and international best practices.	• MOES staff • Unions and other professional associations • Regional MOES and LGA staff • School Directors and teachers • General public	• Structure and guidelines for monitoring formal adopted by MOES • Rigor and quality of self evaluation report of departments • External audit by the MOES	•Agreeing on objective indicators of quality •Fear of retribution for reporting low quality and performance levels	• Staff and organizational units' appraisal guidelines approved and adopted by 2006. • 50% of supervisors and HOD's adopt the new SA and QA by 2008 • 100% 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 plutot 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 formatter 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 deteminer 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 sons censés.

TF-IDF

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 :