

Extraction des concepts-clés à partir du fonds Charcot

Approche *PatternRank*

Ljudmila PETKOVIC^{1,2,3}

`prenom.nom@sorbonne-universite.fr`

¹ Sorbonne Université, Faculté des Lettres, UFR Littératures françaises et comparée, ED 3

² Centre d'étude de la langue et des littératures françaises (CELLF), UMR 8599

³ Observatoire des textes, des idées et des corpus (ObTIC)

Atelier ObTIC

DataLab, BNF

Paris, le 30 avril 2024



Plan

1. Contexte de recherche
2. Extraction des phrases-clés : état de l'art
3. Méthode keybert
4. Méthode *PatternRank*

1. Contexte de recherche

2. Extraction des phrases-clés : état de l'art

3. Méthode keybert

4. Méthode *PatternRank*

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)

JEAN-MARTIN CHARCOT (1825-1893)



Source : [Wikipedia](#).

- père de la neurologie moderne en France au XIX^e s.
 - leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris
- « Mecque de la neurologie »

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)

JEAN-MARTIN CHARCOT (1825-1893)



Source : [Wikipedia](#).

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris

« Mecque de la neurologie »

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)

JEAN-MARTIN CHARCOT (1825-1893)



Source : [Wikipedia](#).

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris

« Mecque de la neurologie »

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)



Source : [Wikipedia](#).

JEAN-MARTIN CHARCOT (1825-1893)

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris
« Mecque de la neurologie »

● Contributions majeures :

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)



Source : [Wikipedia](#).

JEAN-MARTIN CHARCOT (1825-1893)

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris

« Mecque de la neurologie »

- Contributions majeures :
hystérie ← lésion dynamique des circuits cérébraux

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)



Source : [Wikipedia](#).

JEAN-MARTIN CHARCOT (1825-1893)

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris

« Mecque de la neurologie »

- Contributions majeures :
 - hystérie ← lésion dynamique des circuits cérébraux
 - hypnose analyse et traitement des symptômes hystériques

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)



Source : [Wikipedia](#).

JEAN-MARTIN CHARCOT (1825-1893)

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris

« Mecque de la neurologie »

● Contributions majeures :

hystérie	← lésion dynamique des circuits cérébraux
hypnose	analyse et traitement des symptômes hystériques
SEP	description de la <i>sclérose en plaques</i> disséminée

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)



Source : [Wikipedia](#).

JEAN-MARTIN CHARCOT (1825-1893)

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris

« Mecque de la neurologie »

● Contributions majeures :

hystérie	← lésion dynamique des circuits cérébraux
hypnose	analyse et traitement des symptômes hystériques
SEP	description de la <i>sclérose en plaques</i> disséminée
SLA	description de la <i>sclérose latérale amyotrophique</i>

(CAMARGO et al., 2024)

« Napoléon des névroses » ou « Paganini de l'hystérie » (MARMION, 2015)



Source : [Wikipedia](#).

JEAN-MARTIN CHARCOT (1825-1893)

- père de la neurologie moderne en France au XIX^e s.
- leçons cliniques du mardi à l'hôpital de la Salpêtrière à Paris

« Mecque de la neurologie »

- Contributions majeures :

hystérie

← lésion dynamique des circuits cérébraux

hypnose

analyse et traitement des symptômes hystériques

SEP

description de la *sclérose en plaques* disséminée

SLA

description de la sclérose latérale amyotrophique

maladie de Parkinson

concepteur du terme (avec A. Vulpian)

(CAMARGO et al., 2024)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves « réseau scientifique »

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD 1856-1939 théorie psychanalytique

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette
Joseph BABINSKI	1857-1904	pithiatisme, signe de Babinski

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette
Joseph BABINSKI	1857-1904	pithiatisme, signe de Babinski

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette
Joseph BABINSKI	1857-1904	pithiatisme, signe de Babinski

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette
Joseph BABINSKI	1857-1904	pithiatisme, signe de Babinski

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette
Joseph BABINSKI	1857-1904	pithiatisme, signe de Babinski

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

Émile ZOLA 1840–1902 *Lourdes*

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette
Joseph BABINSKI	1857-1904	pithiatisme, signe de Babinski

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

Émile ZOLA	1840–1902	<i>Lourdes</i>
Léon TOLSTOÏ	1828–1910	<i>La Sonate à Kreutzer</i>

(KOEHLER, 2013)

Impact de Charcot sur sa discipline et au-delà

Collaborateurs et élèves

« réseau scientifique »

Sigmund FREUD	1856-1939	théorie psychanalytique
Gilles DE LA TOURETTE	1857-1932	syndrome de Tourette
Joseph BABINSKI	1857-1904	pithiatisme, signe de Babinski

(BROUSSOLLE et al., 2012)

Écrivains naturalistes français et européens

- références à Charcot et aux descriptions de crises hystériques

Émile ZOLA	1840–1902	<i>Lourdes</i>
Léon TOLSTOÏ	1828–1910	<i>La Sonate à Kreutzer</i>
Luigi CAPUANA	1839–1915	<i>La Torture</i>

(KOEHLER, 2013)

Question de recherche

Comment mesurer le degré d'intertextualité entre Charcot et son réseau scientifique au prisme du numérique?

1. Contexte de recherche

2. Extraction des phrases-clés : état de l'art

3. Méthode keybert

4. Méthode *PatternRank*

Définitions de la tâche

Extraction de « **phrases-clés** » (angl. *keyphrases*)

- séquences de plusieurs mots (ex. *sclérose latérale amyotrophique*)
- reflètent plus précisément le contexte sémantique du texte
≠ mots-clés : unigrammes de mot, ex. *sclérose*

Extraction
d'un ensemble de phrases les plus pertinentes à partir d'un texte.

(SCHOPF et al., 2022)

Sélection
Prédiction
des phrases-clés qui résument parfaitement un document donné.

Génération
(XIE et al., 2023)

Définitions de la tâche

Extraction de « **phrases-clés** » (angl. *keyphrases*)

- séquences de plusieurs mots (ex. *sclérose latérale amyotrophique*)
- reflètent plus précisément le contexte sémantique du texte
≠ mots-clés : unigrammes de mot, ex. *sclérose*

Extraction
d'un ensemble de phrases les plus pertinentes à partir d'un texte.

(SCHOPF et al., 2022)

Sélection
des phrases-clés qui résument parfaitement un document donné.

(XIE et al., 2023)

Définitions de la tâche

Extraction de « **phrases-clés** » (angl. *keyphrases*)

- séquences de plusieurs mots (ex. *sclérose latérale amyotrophique*)
- reflètent plus précisément le contexte sémantique du texte
≠ mots-clés : unigrammes de mot, ex. *sclérose*

Extraction
d'un ensemble de phrases les plus pertinentes à partir d'un texte.

(SCHOPF et al., 2022)

Sélection
des phrases-clés qui résument parfaitement un document donné.

(XIE et al., 2023)

Définitions de la tâche

Extraction de « **phrases-clés** » (angl. *keyphrases*)

- séquences de plusieurs mots (ex. *sclérose latérale amyotrophique*)
- reflètent plus précisément le contexte sémantique du texte
≠ mots-clés : unigrammes de mot, ex. *sclérose*

Extraction

d'un ensemble de phrases les plus pertinentes à partir d'un texte.

Sélection

(SCHOPF et al., 2022)

Prédiction

des phrases-clés qui résument parfaitement un document donné.

Génération

(XIE et al., 2023)

Définitions de la tâche

Extraction de « **phrases-clés** » (angl. *keyphrases*)

- séquences de plusieurs mots (ex. *sclérose latérale amyotrophique*)
- reflètent plus précisément le contexte sémantique du texte
≠ mots-clés : unigrammes de mot, ex. *sclérose*

Extraction

Sélection

d'un ensemble de phrases les plus pertinentes à partir d'un texte.

(SCHOPF et al., 2022)

Prédiction

Génération

des phrases-clés qui résument parfaitement un document donné.

(XIE et al., 2023)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots ;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots ;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots ;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches classiques

(GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

GRAPHES

Chaque nœud = mot / groupe de mots ;
chaque arc = probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches sémantiques

(GARAUD, 2022b)

PLONGEMENTS DE MOTS

Représentent l'ensemble des mots d'un vocabulaire sous forme de vecteurs. Distance entre ces vecteurs → mots sémantiquement proches.

- `fastTextRank`¹

PLONGEMENTS CONTEXTUELS

Basés sur les modèles de langue pré-entraînés.
Gèrent mieux des cas ambigus (homographes).

- Key2Vec (MAHATA et al., 2018)
- KeyBERT (GROOTENDORST, 2020)

¹ <https://github.com/jeskim/fasttextrank>

Approches sémantiques

(GARAUD, 2022b)

PLONGEMENTS DE MOTS

Représentent l'ensemble des mots d'un vocabulaire sous forme de vecteurs. Distance entre ces vecteurs → mots sémantiquement proches.

- fastTextRank¹

PLONGEMENTS CONTEXTUELS

Basés sur les modèles de langue pré-entraînés.
Gèrent mieux des cas ambigus (homographes).

- Key2Vec

(MAHATA et al., 2018)

- KeyBERT

(GROOTENDORST, 2020)

1. <https://github.com/jeekim/fasttextrank>

Approches sémantiques

(GARAUD, 2022b)

PLONGEMENTS DE MOTS

Représentent l'ensemble des mots d'un vocabulaire sous forme de vecteurs. Distance entre ces vecteurs → mots sémantiquement proches.

- fastTextRank¹

PLONGEMENTS CONTEXTUELS

Basés sur les modèles de langue pré-entraînés.
Gèrent mieux des cas ambigus (homographes).

- Key2Vec

(MAHATA et al., 2018)

- KeyBERT

(GROOTENDORST, 2020)

1. <https://github.com/jeekim/fasttextrank>

Approches sémantiques

(GARAUD, 2022b)

PLONGEMENTS DE MOTS

Représentent l'ensemble des mots d'un vocabulaire sous forme de vecteurs. Distance entre ces vecteurs → mots sémantiquement proches.

- fastTextRank¹

PLONGEMENTS CONTEXTUELS

Basés sur les modèles de langue pré-entraînés.
Gèrent mieux des cas ambigus (homographes).

- Key2Vec (MAHATA et al., 2018)
- KeyBERT (GROOTENDORST, 2020)

1. <https://github.com/jeekim/fasttextrank>

Approches sémantiques

(GARAUD, 2022b)

PLONGEMENTS DE MOTS

Représentent l'ensemble des mots d'un vocabulaire sous forme de vecteurs. Distance entre ces vecteurs → mots sémantiquement proches.

- fastTextRank¹

PLONGEMENTS CONTEXTUELS

Basés sur les modèles de langue pré-entraînés.
Gèrent mieux des cas ambigus (homographes).

- Key2Vec (MAHATA et al., 2018)
- KeyBERT (GROOTENDORST, 2020)

1. <https://github.com/jeekim/fasttextrank>

1. Contexte de recherche

2. Extraction des phrases-clés : état de l'art

3. Méthode **keybert**

4. Méthode *PatternRank*

Fonctionnement de la librairie keybert

- 1 entrée : un document
- 2 tokénisation du document en mots/phrases-clés candidates
- 3 génération des plongements du document et des mots/phrases-clés
- 4 calcul de la similarité cosinus document : mots/phrases-clés

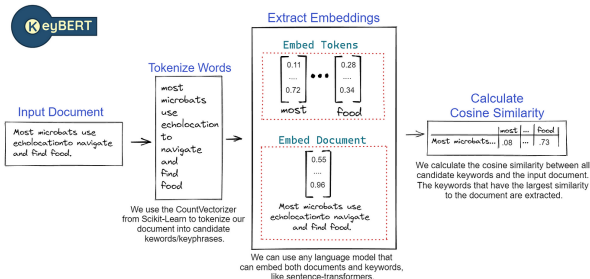


Fig. 1 – Pipeline de la méthode keybert (GROOTENDORST, 2020).

Fonctionnement de la librairie keybert

- 1 entrée : un document
- 2 tokénisation du document en mots/phrases-clés candidates
- 3 génération des plongements du document et des mots/phrases-clés
- 4 calcul de la similarité cosinus document : mots/phrases-clés

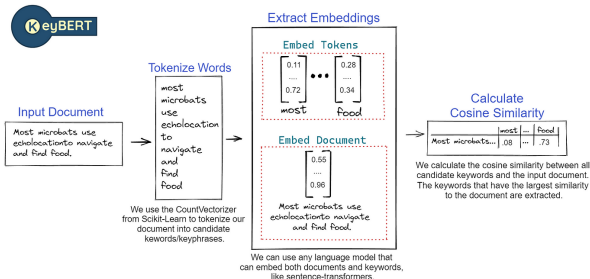


Fig. 1 – Pipeline de la méthode keybert (GROOTENDORST, 2020).

Fonctionnement de la librairie keybert

- 1 entrée : un document
- 2 tokénisation du document en mots/phrases-clés candidates
- 3 génération des plongements du document et des mots/phrases-clés
- 4 calcul de la similarité cosinus document : mots/phrases-clés

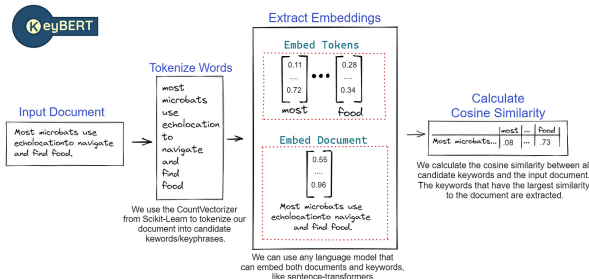


Fig. 1 – Pipeline de la méthode keybert (GROOTENDORST, 2020).

Fonctionnement de la librairie keybert

- 1 entrée : un document
- 2 tokénisation du document en mots/phrases-clés candidates
- 3 génération des plongements du document et des mots/phrases-clés
- 4 calcul de la similarité cosinus document : mots/phrases-clés

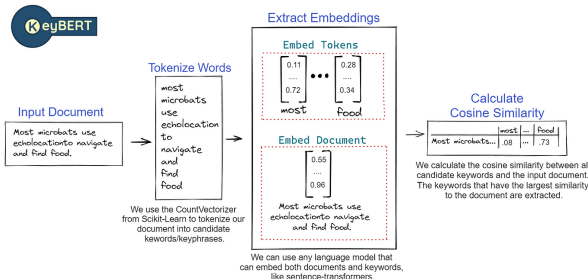


Fig. 1 – Pipeline de la méthode keybert (GROOTENDORST, 2020).

Liste des phrases-clés extraites avec keybert

PHRASE-CLÉ	SCORE
postérieure cordon postérieur	0.5078
scientifique les planches	0.4944
antérieure corne postérieure	0.486
cervico dorsale	0.465
cervicale la figure	0.4311
faisceau postérieur tumeur	0.4276
avoisinante cellule ganglionnaire	0.3836
leucocytes substance granuleuse	0.3474
moëlle épinière 45	0.3381
complètement détruite	0.334

Table 1 – Liste des dix phrases-clés les plus pertinentes selon keybert.

1. Contexte de recherche

2. Extraction des phrases-clés : état de l'art

3. Méthode keybert

4. Méthode *PatternRank*

keybert amélioré – PatternRank

KeyBERT + Keyphrase-Vectorizers = **PatternRank** (SCHOPF et al., 2022)

- extraction des phrases-clés les plus similaires à un document
- préservation de leur grammaticalité grâce aux motifs POS

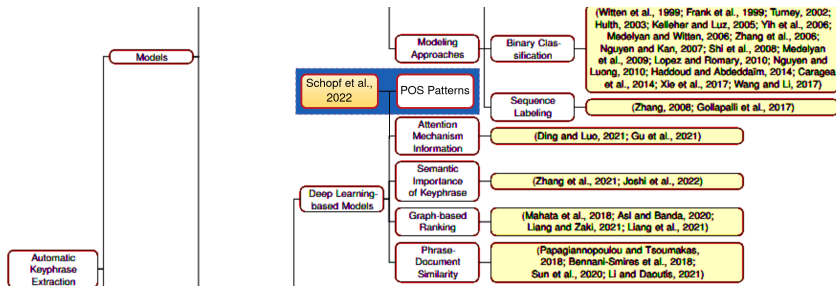


Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

keybert amélioré – PatternRank

KeyBERT + Keyphrase-Vectorizers = **PatternRank** (SCHOPF et al., 2022)

- extraction des phrases-clés les plus similaires à un document
- préservation de leur grammaticalité grâce aux motifs POS

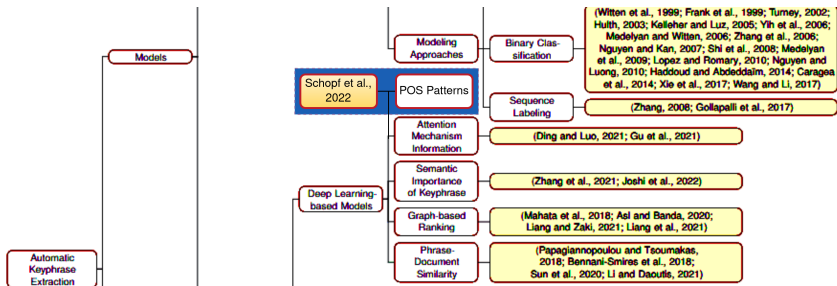


Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

keybert amélioré – *PatternRank*

KeyBERT + Keyphrase-Vectorizers = ***PatternRank*** (SCHOPF et al., 2022)

- extraction des phrases-clés les plus similaires à un document
- préservation de leur grammaticalité grâce aux motifs POS

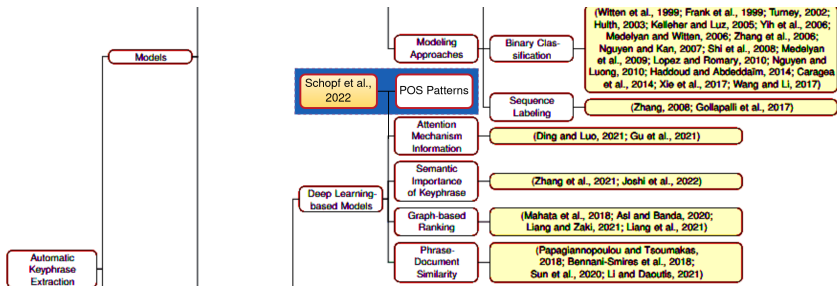


Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

Fonctionnement de la méthode *PatternRank*

- 1 entrée : un seul document texte tokenisé
- 2 étiquetage des tokens avec les balises POS
- 3 sélection des phrases-clés candidates correspondant au modèle POS
- 4 génération des plongements du document et des phrases-clés candidates par un modèle de langue
- 5 calcul des similarités cosinus entre les plongements du document et des phrases-clés candidates + classement des phrases-clés
- 6 extraction des N phrases-clés les plus représentatives

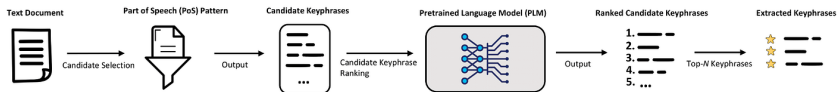


Fig. 3 – Workflow de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

- 1 entrée : un seul document texte tokenisé
- 2 étiquetage des tokens avec les balises POS
- 3 sélection des phrases-clés candidates correspondant au modèle POS
- 4 génération des plongements du document et des phrases-clés candidates par un modèle de langue
- 5 calcul des similarités cosinus entre les plongements du document et des phrases-clés candidates + classement des phrases-clés
- 6 extraction des N phrases-clés les plus représentatives

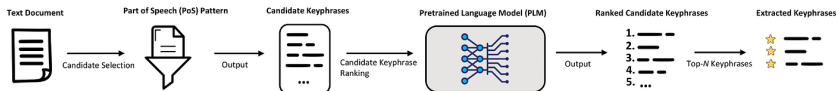


Fig. 3 – Workflow de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

- 1 entrée : un seul document texte tokenisé
- 2 étiquetage des tokens avec les balises POS
- 3 sélection des phrases-clés candidates correspondant au modèle POS
- 4 génération des plongements du document et des phrases-clés candidates par un modèle de langue
- 5 calcul des similarités cosinus entre les plongements du document et des phrases-clés candidates + classement des phrases-clés
- 6 extraction des N phrases-clés les plus représentatives

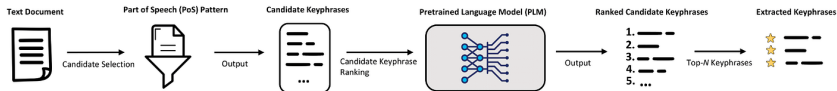


Fig. 3 – Workflow de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

- 1 entrée : un seul document texte tokenisé
- 2 étiquetage des tokens avec les balises POS
- 3 sélection des phrases-clés candidates correspondant au modèle POS
- 4 génération des plongements du document et des phrases-clés candidates par un modèle de langue
- 5 calcul des similarités cosinus entre les plongements du document et des phrases-clés candidates + classement des phrases-clés
- 6 extraction des N phrases-clés les plus représentatives

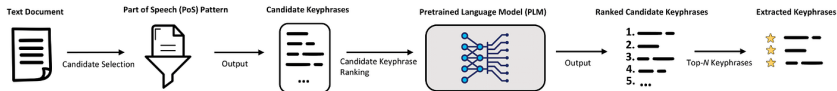


Fig. 3 – Workflow de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

- 1 entrée : un seul document texte tokenisé
- 2 étiquetage des tokens avec les balises POS
- 3 sélection des phrases-clés candidates correspondant au modèle POS
- 4 génération des plongements du document et des phrases-clés candidates par un modèle de langue
- 5 calcul des similarités cosinus entre les plongements du document et des phrases-clés candidates + classement des phrases-clés
- 6 extraction des N phrases-clés les plus représentatives

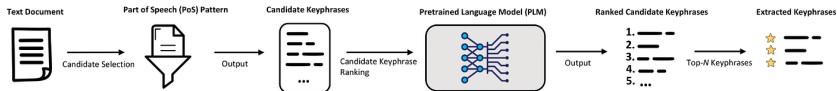


Fig. 3 – Workflow de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

- 1 entrée : un seul document texte tokenisé
- 2 étiquetage des tokens avec les balises POS
- 3 sélection des phrases-clés candidates correspondant au modèle POS
- 4 génération des plongements du document et des phrases-clés candidates par un modèle de langue
- 5 calcul des similarités cosinus entre les plongements du document et des phrases-clés candidates + classement des phrases-clés
- 6 extraction des N phrases-clés les plus représentatives

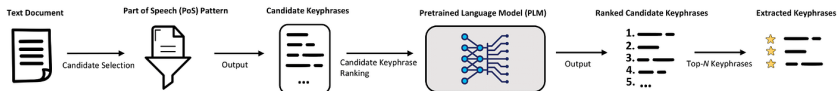


Fig. 3 – Workflow de la méthode *PatternRank* (SCHOPF et al., 2022).

Liste des phrases-clés avec keyphrase-vectorizers

PHRASE-CLÉ	SCORE
paroi épaissie	0.9276
histologie fine	0.9193
tissu gingival	0.9179
21c leçon	0.9162
travaux récents	0.9152
entrecroisement des pyramides	0.9145
érysipèle périodique annuel	0.9135
cicatriciel	0.9118
fibromes	0.9109
affections	0.9091

Table 2 – Liste des dix phrases-clés les plus pertinentes selon keyphrase-vectorizers.

Utilisation des librairies keybert et keyphrase-vectorizers

Ressources en ligne :

- [Lien Google Colab](#)
- pré-requis :
 - bonne connexion internet
 - disposer d'un ordinateur
- [Dépôt GitHub](#)

Utilisation des librairies keybert et keyphrase-vectorizers

Ressources en ligne :

- [Lien Google Colab](#)

pré-requis :

- bonne connexion Internet
- mémoire RAM suffisante
- [Dépôt GitHub](#)

Utilisation des bibliothèques keybert et keyphrase-vectorizers

Ressources en ligne :

- [Lien Google Colab](#)

pré-requis :

- bonne connexion Internet
 - mémoire RAM suffisante
- [Dépôt GitHub](#)

Utilisation des bibliothèques keybert et keyphrase-vectorizers

Ressources en ligne :

- [Lien Google Colab](#)

pré-requis :

- bonne connexion Internet
- mémoire RAM suffisante
- [Dépôt GitHub](#)

Utilisation des librairies keybert et keyphrase-vectorizers

Ressources en ligne :

- [Lien Google Colab](#)

pré-requis :

- bonne connexion Internet
- mémoire RAM suffisante
- [Dépôt GitHub](#)

Passage à l'échelle

Pour traiter de grands corpus, il existe la possibilité de demander l'accès à la plateforme technologique [MESU](#), hébergée par SACADO (Service d'Aide au Calcul et à l'Analyse de Données).

Elle est composée d'un supercalculateur, d'un environnement de virtualisation et d'un système de stockage de données.

Passage à l'échelle

Pour traiter de grands corpus, il existe la possibilité de demander l'accès à la plateforme technologique [MESU](#), hébergée par SACADO (Service d'Aide au Calcul et à l'Analyse de Données).

Elle est composée d'un supercalculateur, d'un environnement de virtualisation et d'un système de stockage de données.

Références I



BOUGOUIN, A., F. BOUDIN et B. DAILLE (2013). TopicRank : Graph-Based Topic Ranking for Keyword Extraction. In : *International Joint Conference on Natural Language Processing (IJCNLP)*, p. 543-551 (*voir pp. 31-38*).



BROUSSOLLE, E., J. POIRIER, F. CLARAC et J.-G. BARBARA (2012). Figures and institutions of the neurological sciences in Paris from 1800 to 1950. Part III : Neurology. In : *Revue Neurologique* 168.4, p. 301-320 (*voir pp. 13-23*).



CAMARGO, C. H. F., L. COUTINHO, Y. CORREA NETO, E. ENGELHARDT, P. MARANHÃO FILHO, O. WALUSINSKI et H. A. G. TEIVE (2024). Jean-Martin Charcot : the polymath. In : *Arquivos de Neuro-psiquiatria* 81, p. 1098-1111 (*voir pp. 4-12*).



CAMPOS, R., V. MANGARAVITE, A. PASQUALI, A. JORGE, C. NUNES et A. JATOWT (2020). YAKE! Keyword extraction from single documents using multiple local features. In : *Information Sciences* 509, p. 257-289 (*voir pp. 31-38*).



GARAUD, D. (22 fév. 2022a). Extraire automatiquement les concepts et mots-clés d'un texte (Part I : Les méthodes dites classiques). Oncrawl. (Visité le 09/04/2024) (*voir pp. 31-38*).



GARAUD, D. (22 fév. 2022b). Extraire automatiquement les concepts et mots-clés d'un texte (Part II : approche sémantique). Oncrawl. (Visité le 09/04/2024) (*voir pp. 39-43*).



GROOTENDORST, M. (2020). KeyBERT : Minimal keyword extraction with BERT. Version v0.3.0 (*voir pp. 39-43, 45-48*).

Références II



KOEHLER, P. J. (2013). Charcot, La Salpêtrière, and Hysteria as Represented in European Literature. In : *Progress in Brain Research* 206, p. 93-122 (voir pp. 13-23).



MAHATA, D., J. KURIAKOSE, R. SHAH et R. ZIMMERMANN (2018). Key2Vec : Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 634-639 (voir pp. 39-43).



MARMION, J.-F. (2015). *Freud et la psychanalyse*. Sciences Humaines (voir pp. 4-12).



MIHALCEA, R. et P. TARAU (2004). TextRank : Bringing Order into Text. In : *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Sous la dir. de D. LIN et D. WU. Barcelona, Spain : Association for Computational Linguistics, p. 404-411 (voir pp. 31-38).



ROSE, S., D. ENGEL, N. CRAMER et W. COWLEY (2010). Automatic Keyword Extraction from Individual Documents. In : *Text Mining : Applications and Theory*, p. 1-20 (voir pp. 31-38).



SCHOPF, T., S. KLIMEK et F. MATTHES (2022). PatternRank : Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In : *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) – KDIR*. INSTICC. SciTePress, p. 243-248 (voir pp. 26-30, 51-59).



SPARCK JONES, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. In : *Journal of documentation* 28.1, p. 11-21 (voir pp. 31-38).

Références III



WAN, X. et J. XIAO (août 2008). CollabRank : Towards a Collaborative Approach to Single-Document Keyphrase Extraction. In : *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Sous la dir. de D. SCOTT et H. USZKOREIT. Manchester, UK : Coling 2008 Organizing Committee, p. 969-976 ([voir pp. 31-38](#)).



XIE, B., J. SONG, L. SHAO, S. WU, X. WEI, B. YANG, H. LIN, J. XIE et J. SU (2023). From Statistical Methods to Deep Learning, Automatic Keyphrase Prediction : A Survey. In : *Information Processing & Management* 60.4, p. 103382 ([voir pp. 26-30, 51-53](#)).