**IA2I**



Projet de l’Analyse de Données

Textuelles : Journal

Encadré par:

* Ouarab Aziz

**Réalisé par:**

* **Chaimaa BOUABD**
* **Aicha AIT ELBAZ**

Contenue du Rapport

[I. Introduction 3](#_Toc1)

[II. Objectif 3](#_Toc2)

[III. Jeux de donnée 3](#_Toc3)

[IV. Importation des donnée 4](#_Toc4)

[V. Prétraitement 4](#_Toc5)

[VI. Analyse a l’aide de L’ AFC 6](#_Toc6)

[VII. Analyse a l’aide de L’ ACP 12](#_Toc7)

[VIII. Classification 17](#_Toc8)

[IX. Conclusion 17](#_Toc9)

## Introduction

Le text mining de nouvelles consiste à extraire des informations utiles des articles de presse et à découvrir des tendances cachées dans les médias. Cela peut être utile pour les entreprises, les gouvernements et les organisations de recherche qui cherchent à comprendre l'opinion publique et à suivre l'actualité.

L’Analyse Textuelle utilise des techniques de traitement du langage naturel et de l'apprentissage automatique pour analyser et comprendre le sens des articles de presse. Cela peut inclure l'identification de mots-clés, la reconnaissance de sentiments et l'analyse de thèmes.

Elle peut être utilisé pour des tâches telles que la classification de documents, la génération de résumés automatiques, la détection de plagiat et la recherche de réponses à des questions.

En utilisant des outils de text mining de nouvelles, les entreprises peuvent mieux comprendre l'opinion publique et les tendances du marché, améliorer leur stratégie de communication et prendre des décisions plus éclairées.

## Objectif

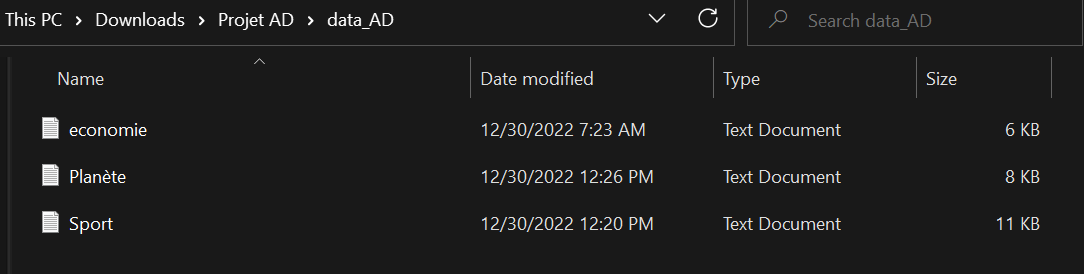
L'objectif de ce projet est d'analyser un texte contenant des données textuelles brutes à l'aide des méthodes que nous avons apprises au cours de ce programme d'analyse de données (ACP,AFC,ACM....). Les objectifs principaux sont les suivants :

* Traiter et présenter les données à analyser sous forme de tableaux appropriés .
* Appliquer une analyse factorielle : Analyse en composantes principales ou Analyse des correspondances suivant la nature des données à analyser.
* Une typologie à l’aide d’une classification.
* Faire toutes les implications des notions vues dans le cours et interprétations utiles à la description des données.

## Jeux de donnée

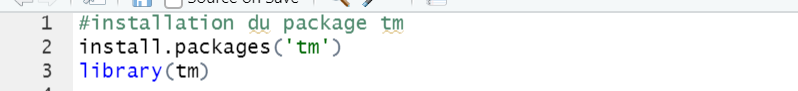
. Le jeux de donnée qu’on va utiliser au cours de ce projet extraire de ces pages web :

* [Football : le champion du monde 2018 Blaise Matuidi annonce la fin de sa carrière (france24.com)](https://www.france24.com/fr/sports/20221223-football-le-champion-du-monde-2018-blaise-matuidi-annonce-la-fin-de-sa-carri%C3%A8re)
* ["La mission est perdue", la fusée européenne Vega-C essuie son premier échec (france24.com)](https://www.france24.com/fr/sciences/20221221-la-mission-est-perdue-la-fus%C3%A9e-europ%C3%A9enne-vega-c-essuie-son-premier-%C3%A9chec)
* [Climat : l'UE réforme son marché carbone avec l'obligation de fortement diminuer les émissions (france24.com)](https://www.france24.com/fr/europe/20221218-l-ue-s-accorde-sur-une-r%C3%A9forme-de-son-march%C3%A9-carbone-pour-atteindre-ses-objectifs-climatiques)
* [L’Union européenne approuve une taxe carbone aux frontières (france24.com)](https://www.france24.com/fr/europe/20221213-l-ue-approuve-une-taxe-carbone-sur-les-importations-de-produits-polluants)



## Importation des donnée

On a besoin d’installer le package “tm ” .il fournit des outils pour lire, prétraiter et analyser des données textuelles, ainsi que pour créer et manipuler divers types de documents textuels.La bibliothèque “tm ”comprend également des fonctions pour calculer diverses statistiques et métriques liées aux données textuelles,



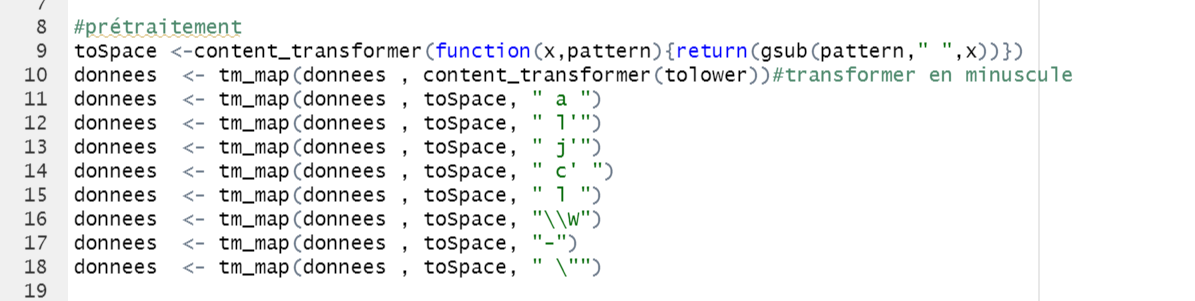
Création du corpus dans variable docs .



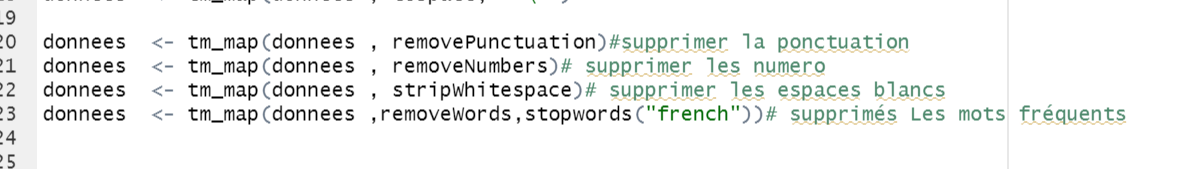
## Prétraitement

Cette partie consiste à nettoyer les donnée :

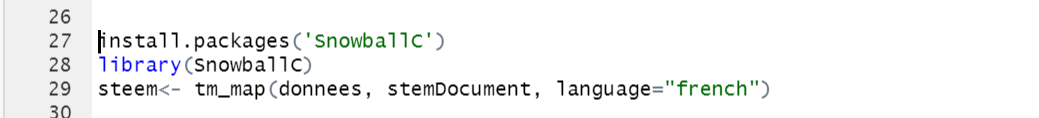
* transformer en minuscule et remplacer ( a , l’, j’,... ) par des espaces.
* Supprimer les caractères spéciaux .



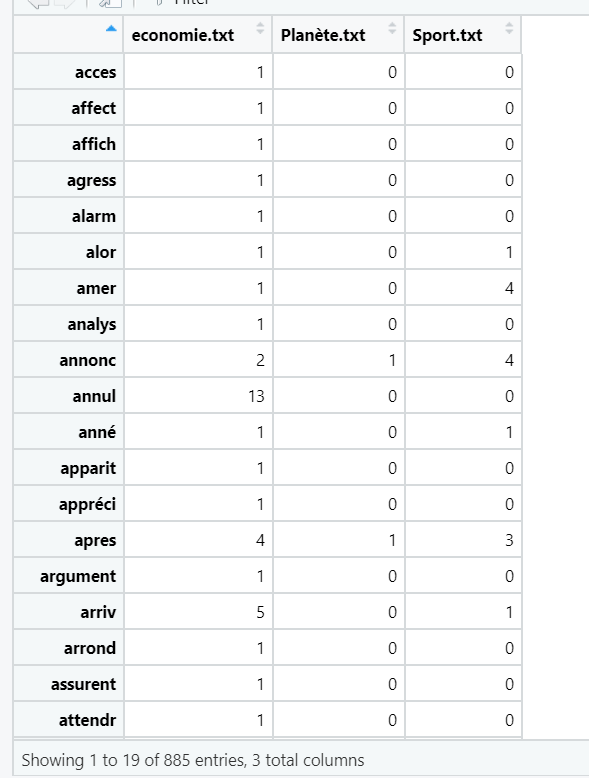
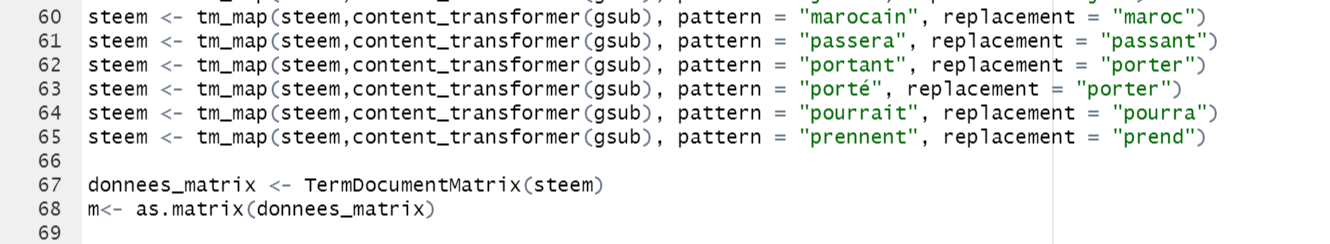
* supprimer la ponctuation
* supprimer les numero
* supprimés Les mots fréquents
* supprimer les espaces blancs



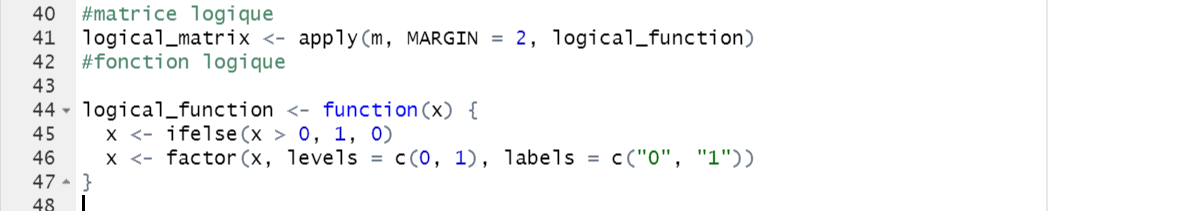
* Pour la partie de concaténation on utilise le package SwonballC pour stemming .



* Et enfin, on rend les données sous forme d'une matrice “m”.



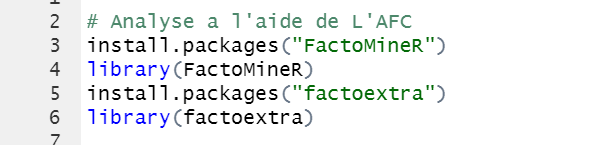
* la matrice logique



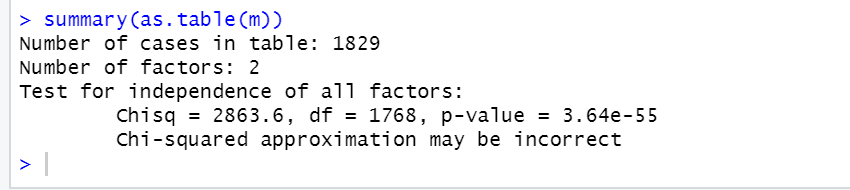
## Analyse a l’aide de L’ AFC

1. Instalation des packge utiliser

* Pour faire l’AFC on va utilser le packages “FactoMineR”



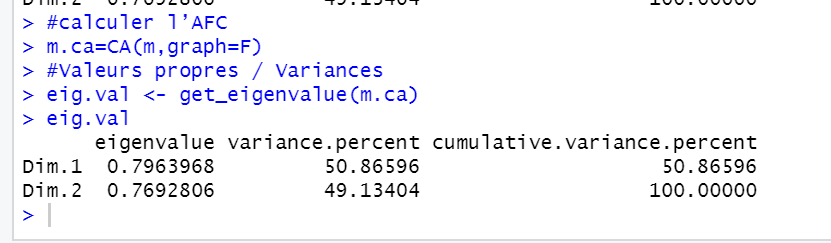
1. Test de chi-2



la valeur p est très petite (p-value = 3.64e-55) Cela signifie que les données observées sont très peu probables d'avoir été observées par hasard seul et que l'hypothèse nulle est très peu probable d'être vraie. En d'autres termes, il y a une très forte preuve en faveur de l'hypothèse alternative et contre l'hypothèse nulle.

Donc les variables de ligne et de colonne sont statistiquement significativement associées

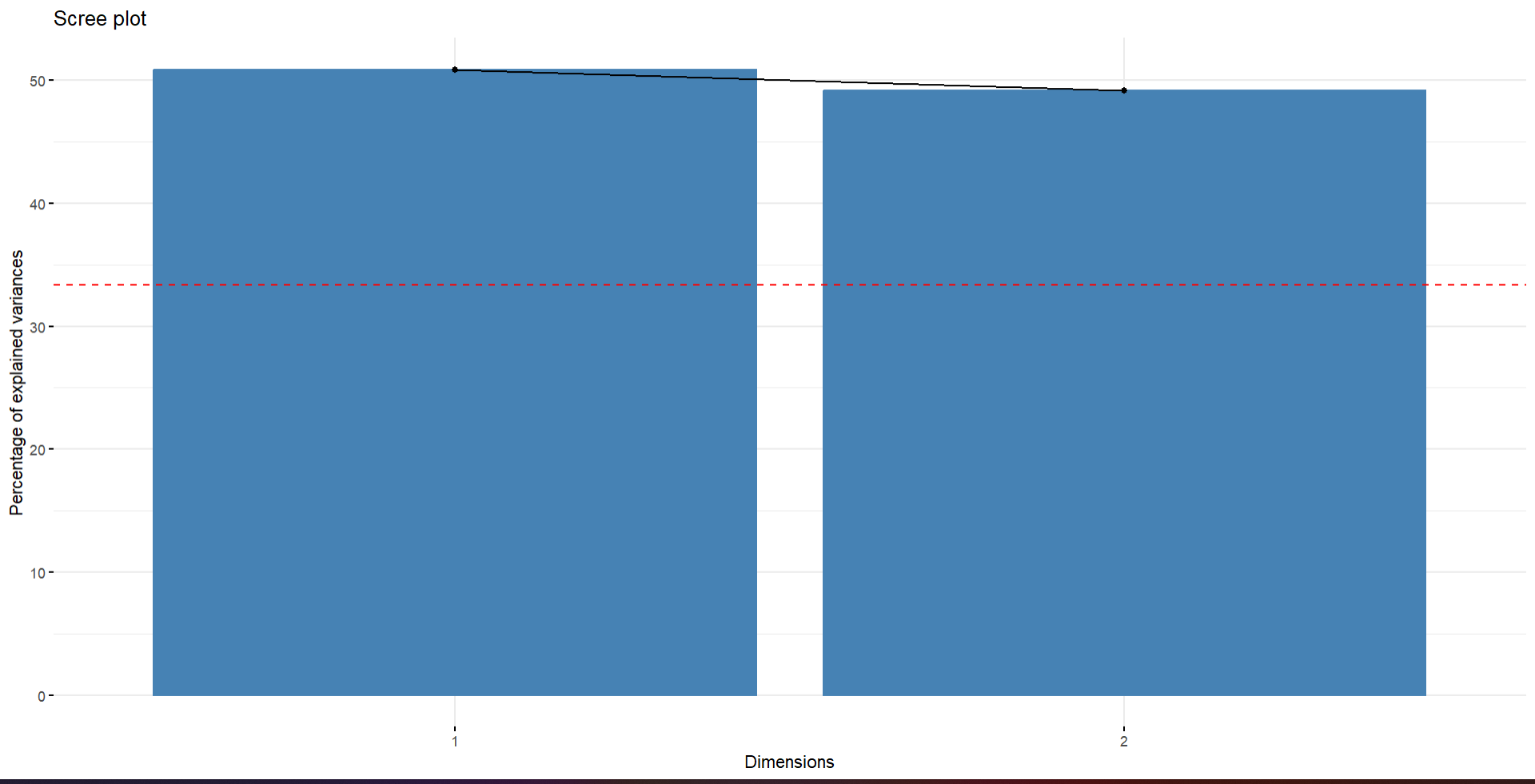
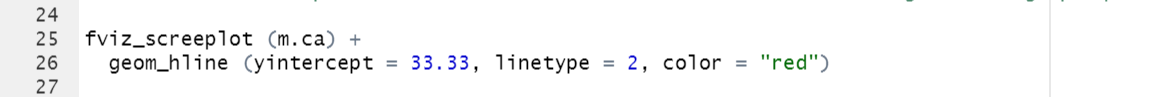
1. Visualisation et interprétation
2. **calculer l’AFC et Valeurs propres / Variances**



#100% de la variance est expliquée par les deux dimensions.

# les deux dimensions représentent la totalité de la variabilité.

Une autre méthode pour déterminer le nombre de dimensions est de regarder le graphique des valeurs propres avec une droite en pointillée rouge spécifiant la valeur propre moyenne

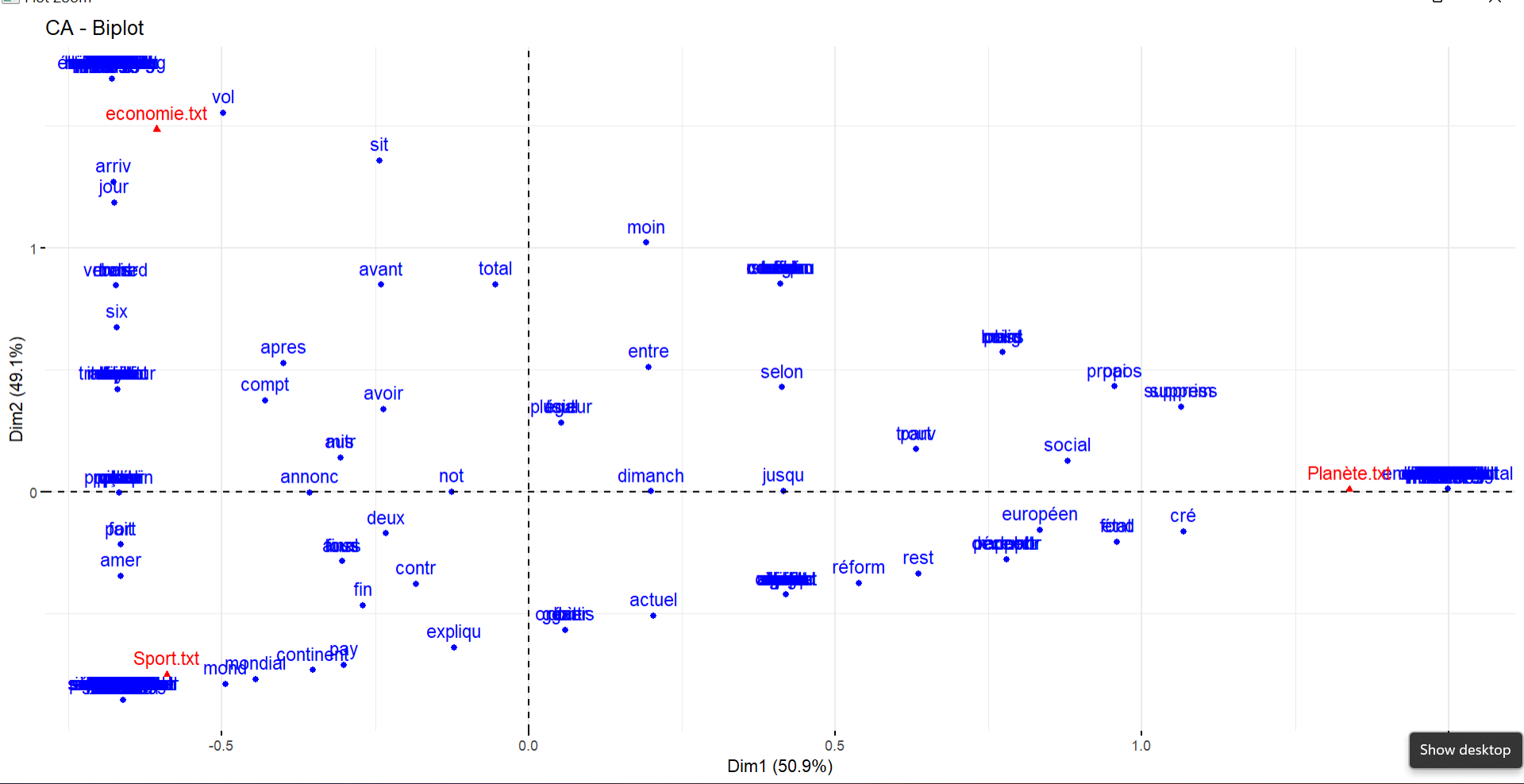


Selon le graphique ci-dessus,les deux dimensions doivent être considérées pour l’interprétation de la solution

1. **Visualisation du biplot**

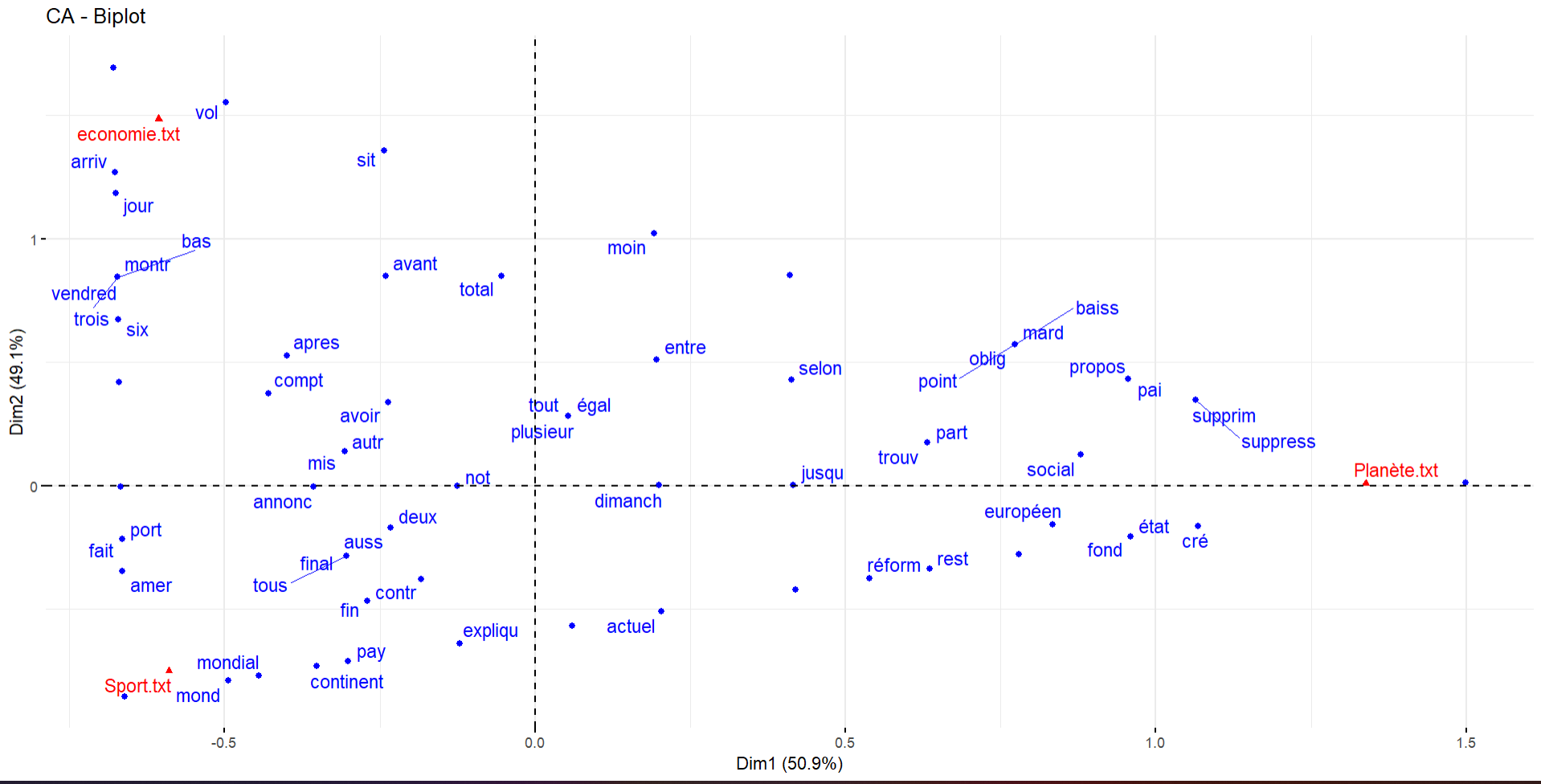
#Visualisation

**fviz\_ca\_biplot (m.ca, repel = F)**



# repel = TRUE pour éviter le chevauchement de texte

**fviz\_ca\_biplot (m.ca, repel = TRUE)**



#interpretation

#Ce graphique montre que:

# les lignes supprim ,suppress et cré sont associées le plus à la colonne planéte.txt

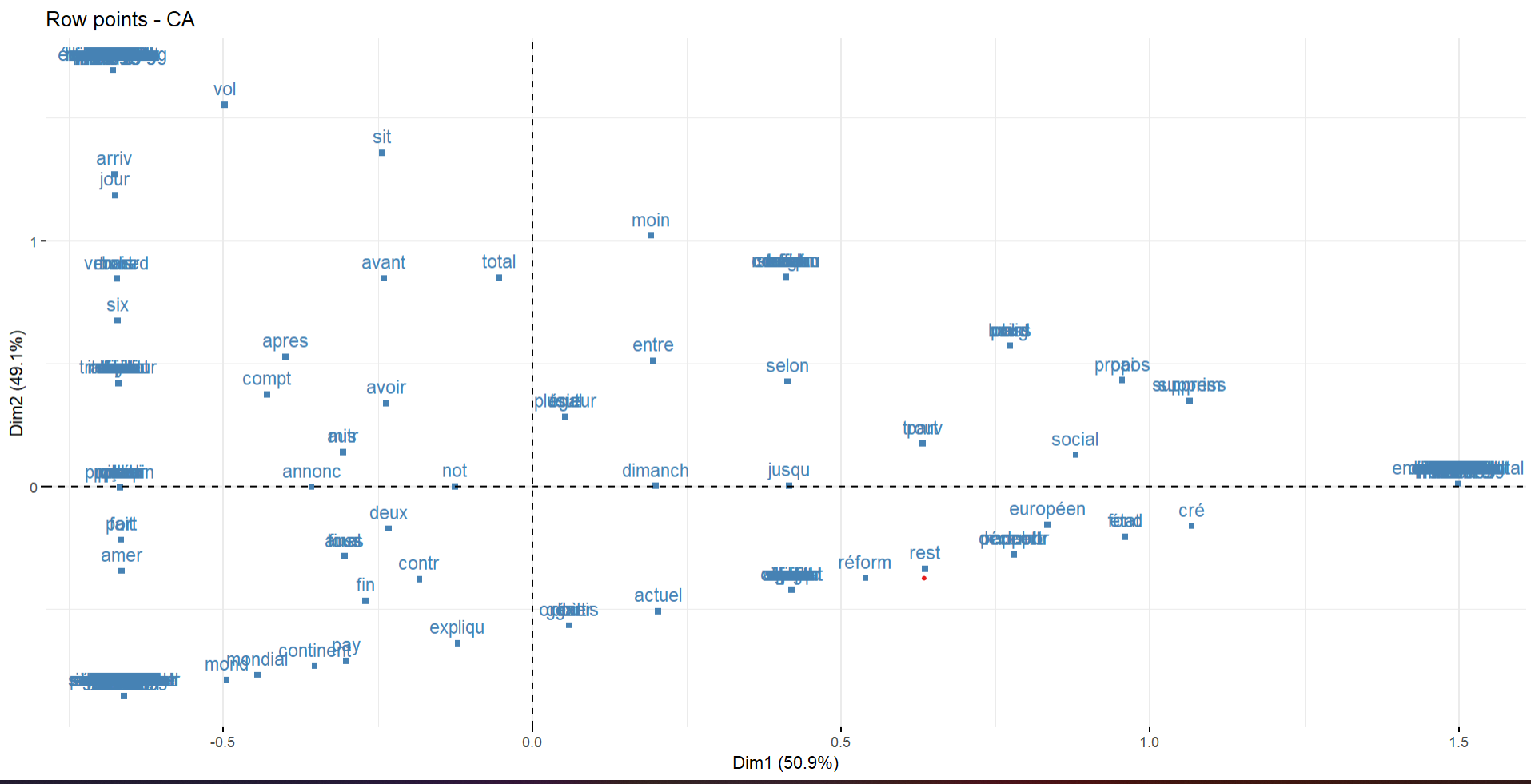
# les lignes mondial,mond, continent sont associées le plus à la colonne sport.txt

# les lignes vol, arriv, jour,bas sont associées le plus à la colonne economie.txt

1. **Graphique des points lignes**

* visualiser uniquement les points lignes a l’aide de la fonction :

fviz\_ca\_row (m.ca, col.row = "steelblue", shape.row = 15)



#Le graphique ci-dessus montre les relations entre les points lignes:

#Les lignes avec un profil similaire sont regroupées.

#Les lignes corrélées négativement sont positionnées sur des côtés opposés de l’origine de du graphique (quadrants opposés).

#La distance entre les points lignes et l’origine mesure la qualité des points lignes sur le graphique. Les points lignes qui sont loin de l’origine sont bien représentés sur le graphique.

1. **Graphique des points lignes**

# Colorer en fonction du cos2:

**fviz\_ca\_row (m.ca, col.row = "cos2",**

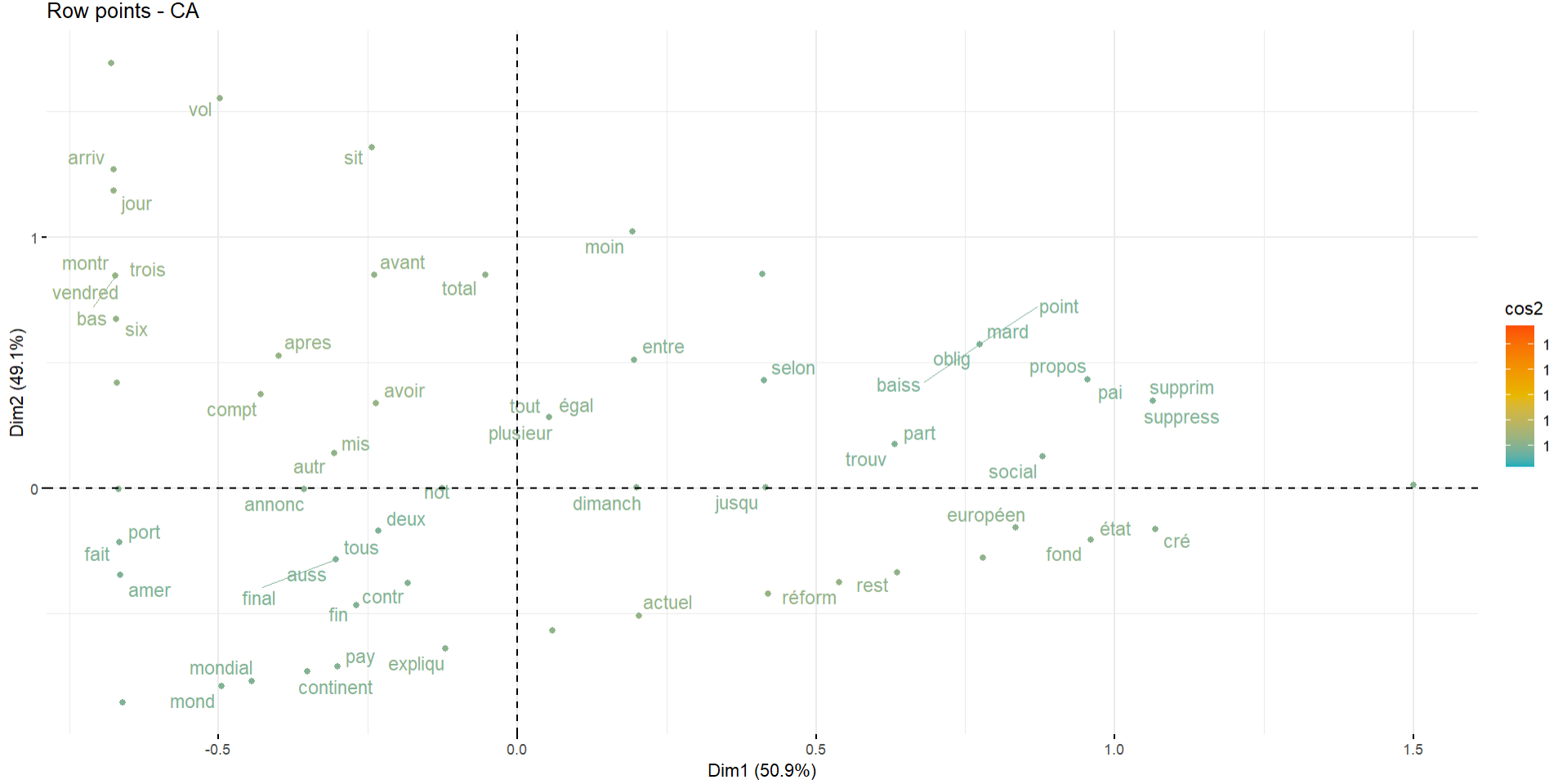
**gradient.cols = c ("#00AFBB", "#E7B800", "#FC4E07"),**

**repel = TRUE)**

#les variables à faible valeur cos2 seront colorées en “white” (blanc)

#les variables avec des valeurs moyennes de cos2 seront colorées en “blue” (bleu)

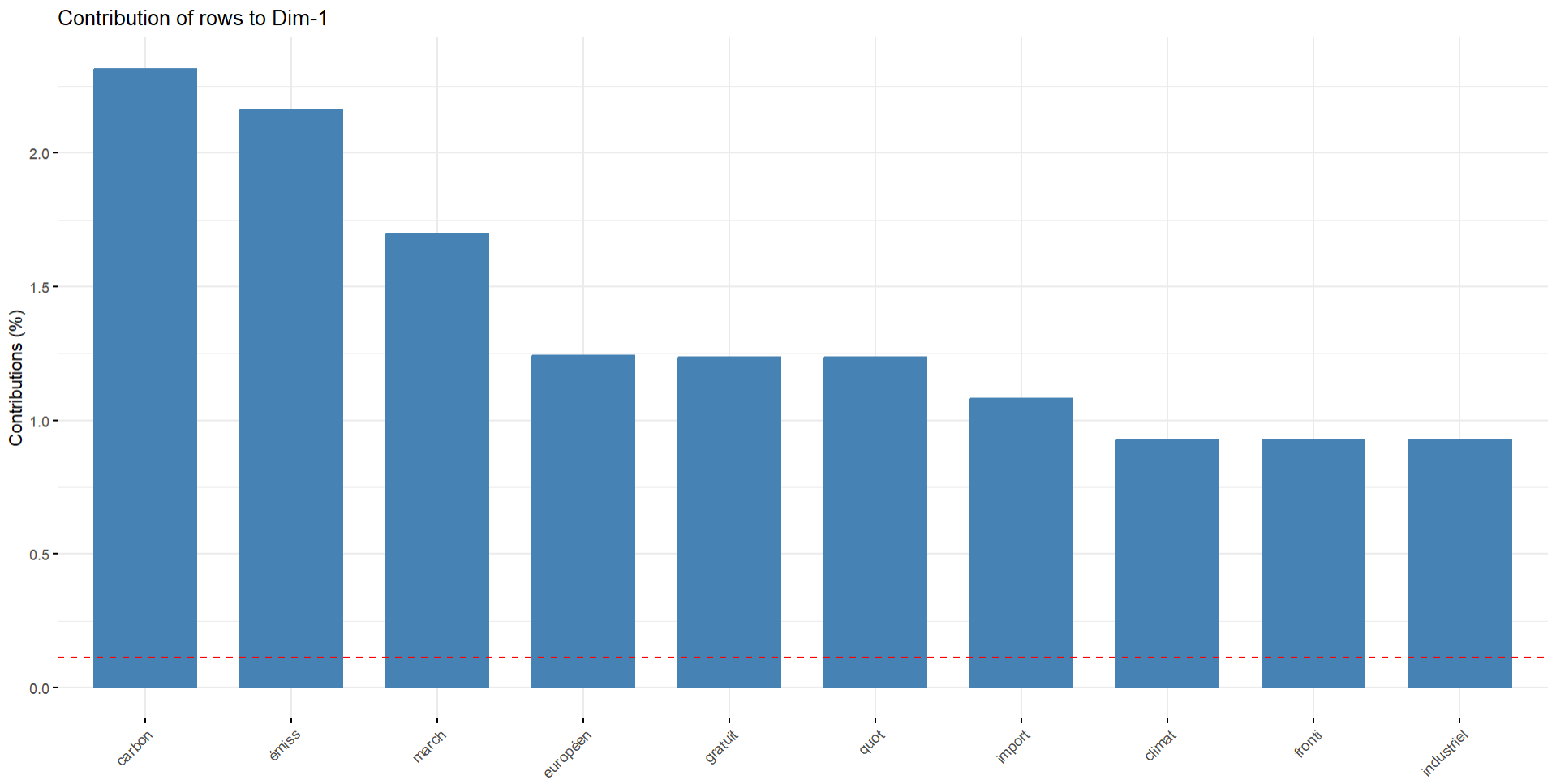
#les variables avec des valeurs élevées de cos2 seront colorées en “red” (rouge)



1. **Contributions des lignes aux dimensions**

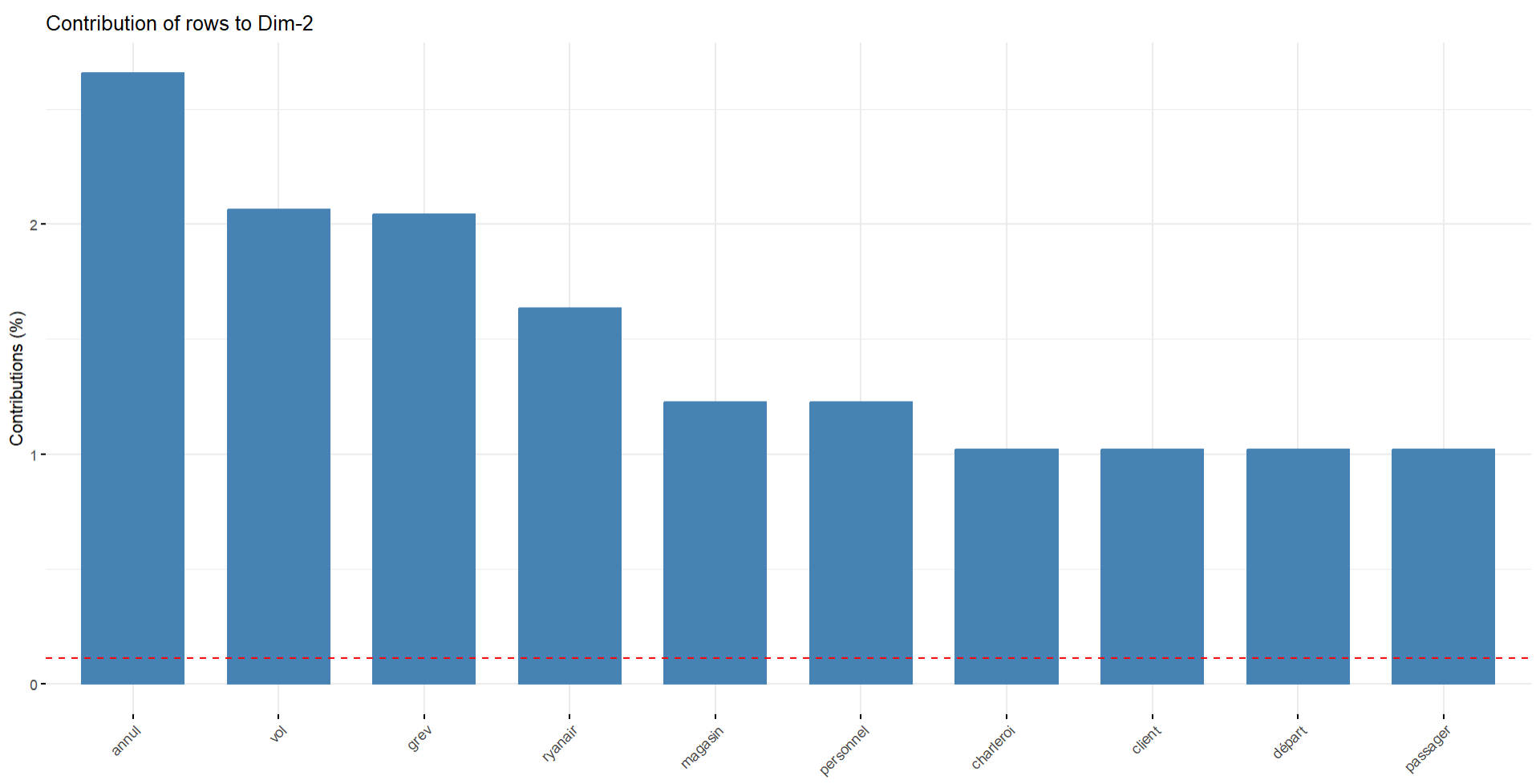
# Contributions des lignes à la dimension 1 pour les top 10 lignes.

**fviz\_contrib(m.ca, choice = "row", axes = 1, top = 10)**



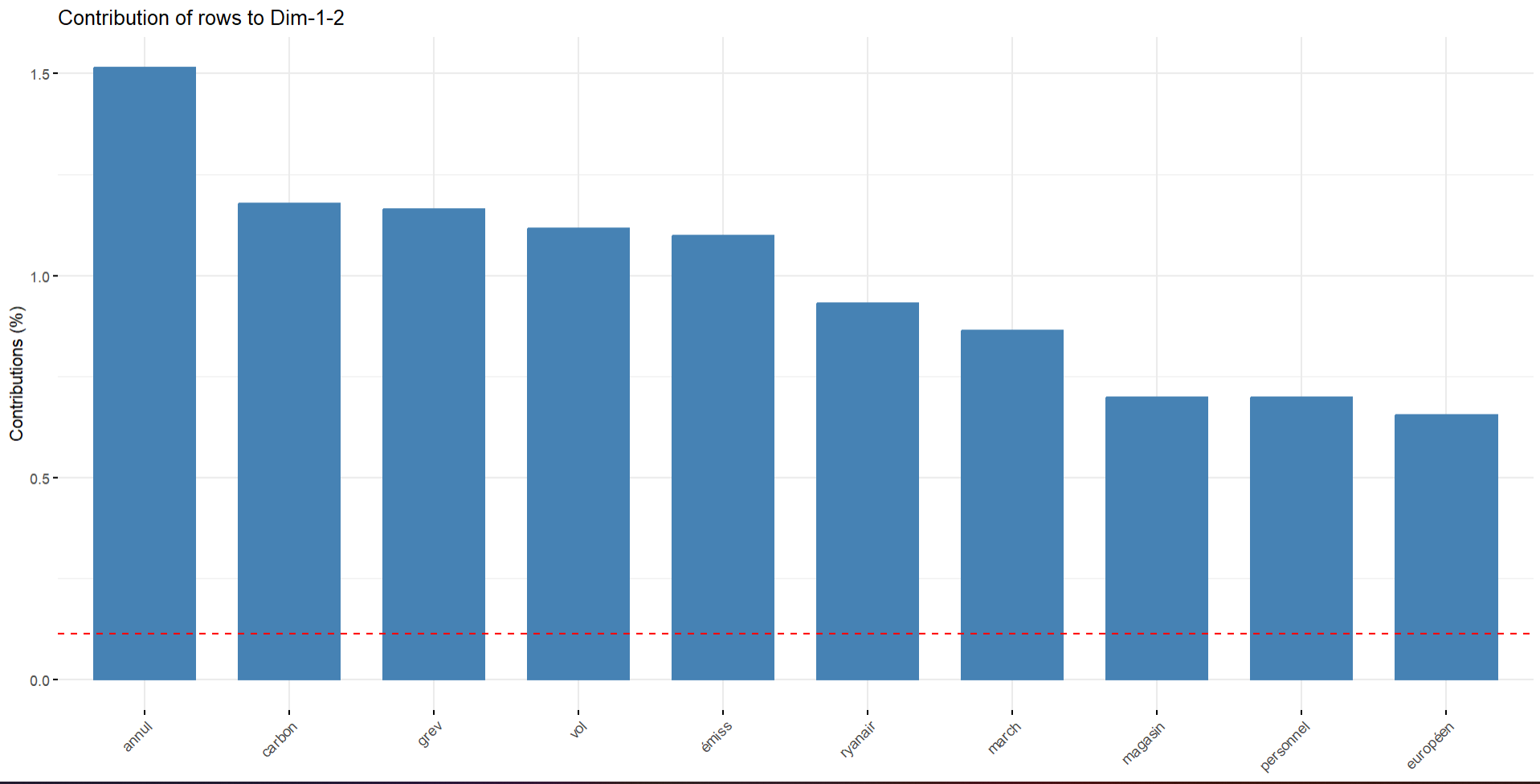
# Contributions des lignes à la dimension 2 pour les top 10 lignes.

**fviz\_contrib(m.ca, choice = "row", axes = 2, top = 10)**



# Contribution totale aux dimensions 1 et 2 pour les top 10 lignes.

**fviz\_contrib (m.ca, choice = "row", axes = 1:2, top = 10)**

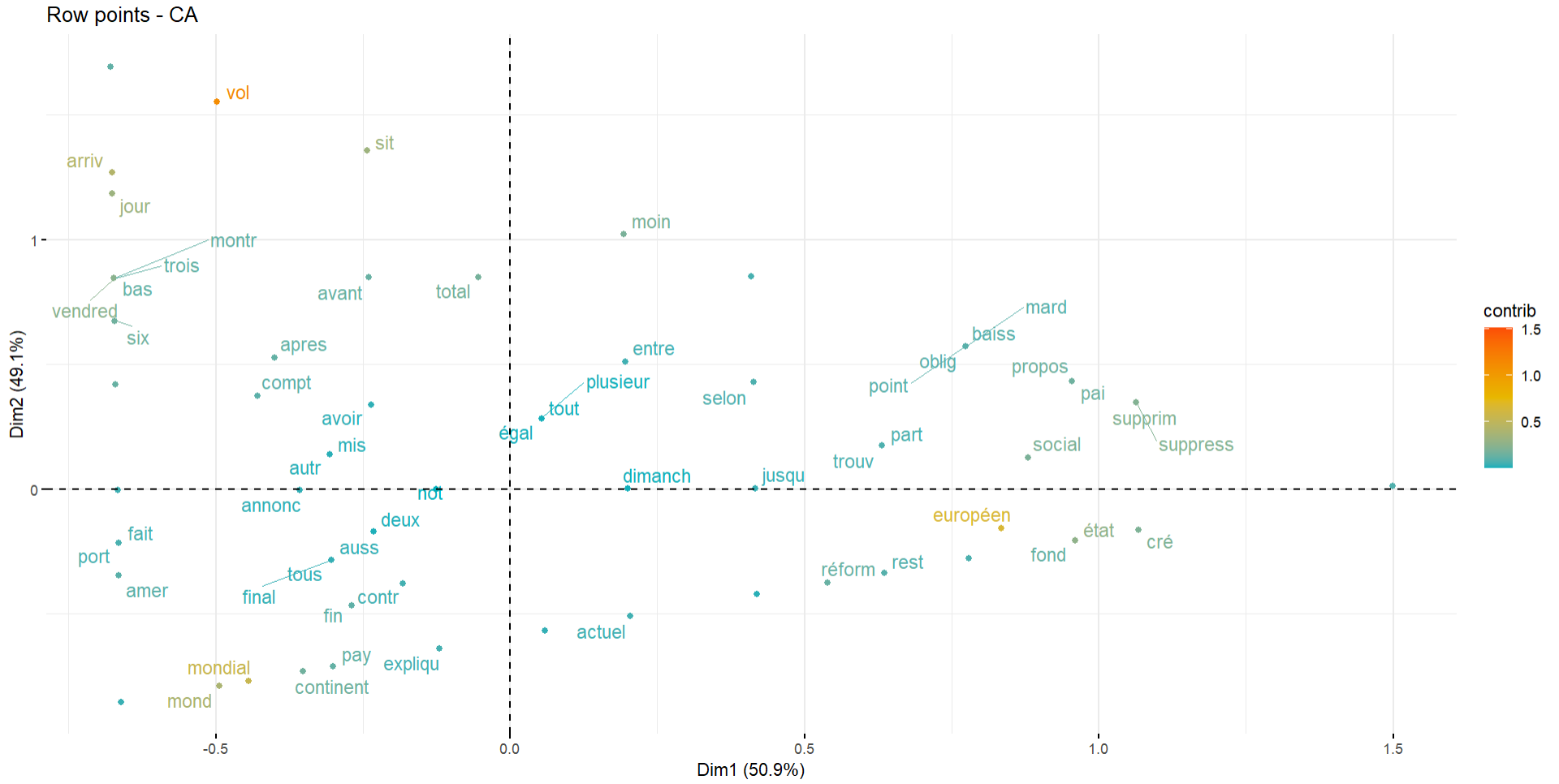


#Les lignes les plus importantes (ou contributives) peuvent être mise en avant sur le graphique comme suit:

**fviz\_ca\_row (m.ca, col.row = "contrib",**

**gradient.cols = c ("#00AFBB", "#E7B800", "#FC4E07"),**

**repel = TRUE)**



#Le graphique donne une idée de la contribution des lignes aux différents pôles des dimensions.

#Selon ce graphe les catégories suppress ,supprim et propos ont une contribution importante

#au pôle positif de la première dimension

#, tandis que les catégories arriv et jour ont une contribution majeure au pôle négatif de

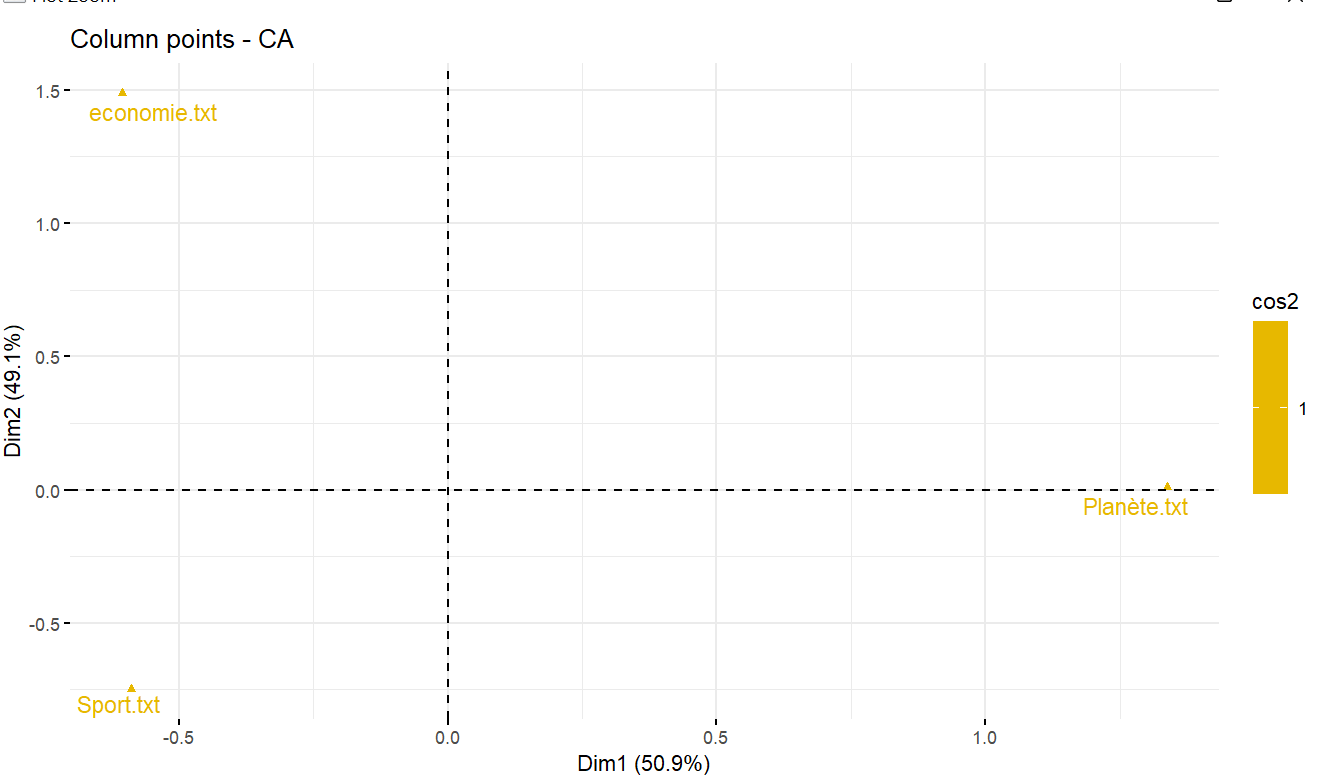
#la première dimension

1. **Qualité de représentation des colonnes**

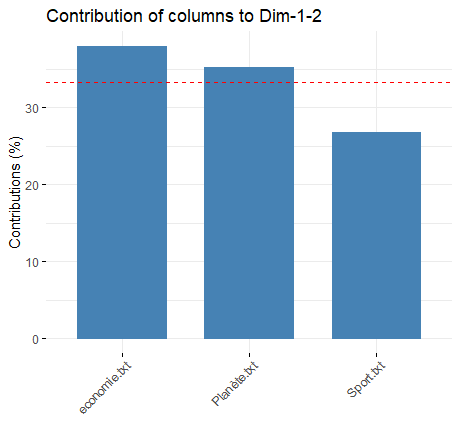
# Colorer en fonction du cos2:

**fviz\_ca\_col (m.ca, col.col = "cos2",**

**gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)**



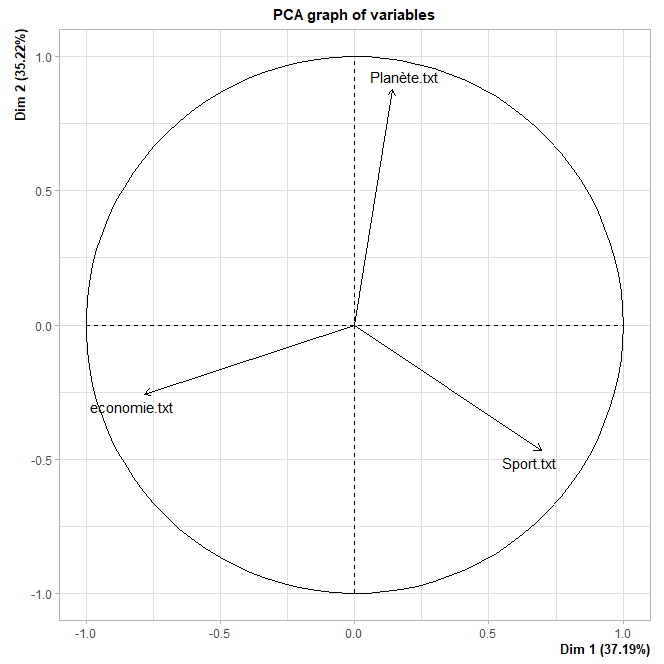
1. **la contribution des colonnes aux deux dimensions**



## Analyse a l’aide de L’ ACP

1. Etude d’ACP

**pca=PCA(m,scale.unit=TRUE,graph=T)**



#intrérprétations de graphe

le cercle de corrélation permet de visualiser la relation existe entre les variables sur un plan,

\* Les flèches sont proches à toucher le cercle =>on peut dire que les variables sont intérprétables sur ce Plan

\* Étudions donc les corrélations entre les variables initiales et les composantes principales

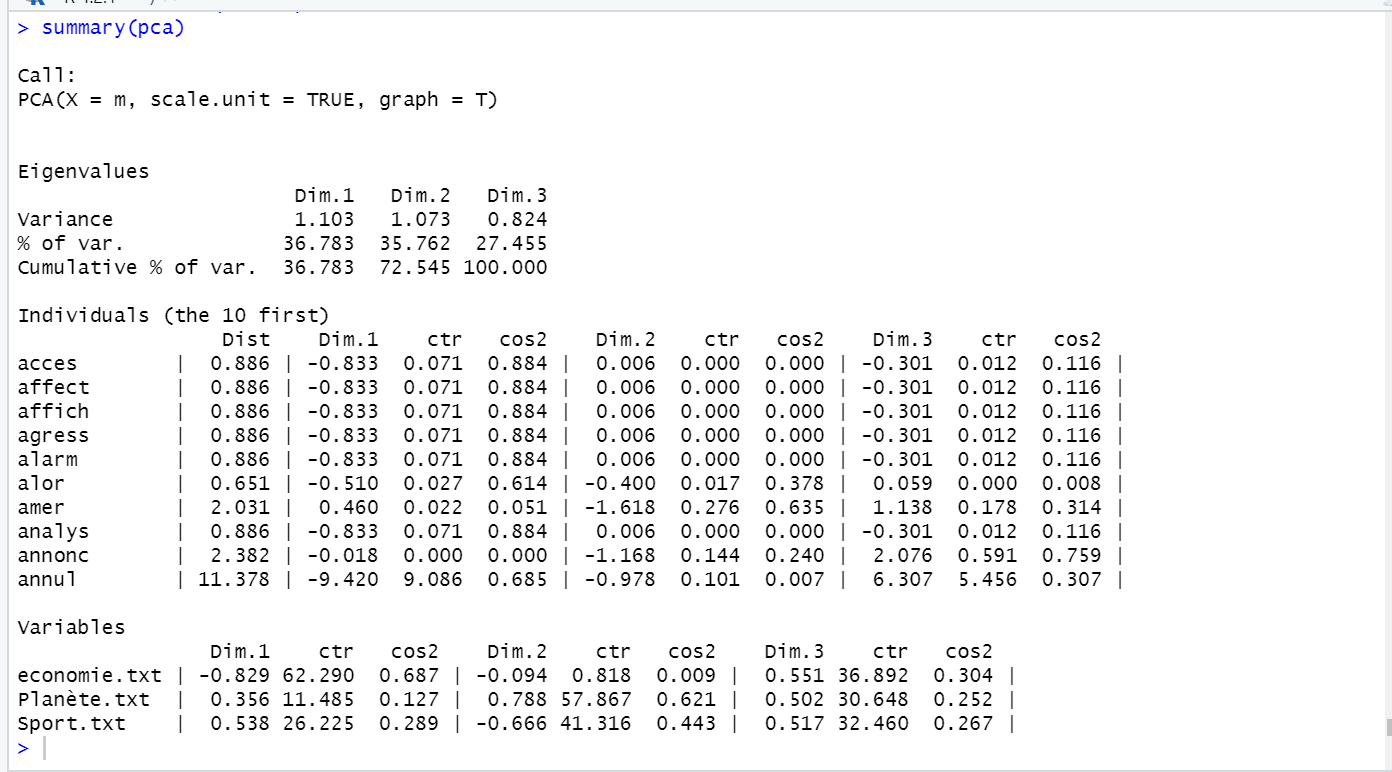
|  |  |  |  |
| --- | --- | --- | --- |
|  | **economie** | **planète** | **sport** |
| **DIM1** | Fortement intercorréler | Trés peu corréler | Fortement corréler |
| **DIM2** | Très peu intercorréller | Fortement corréler | Peu intercorréler |

#on peut lancer summary(pca) pour avoir les informations sur les valeurs propres,variances

#et les pourcentages d'enertie associé à chaque dimension , Ainsi que des informations sur

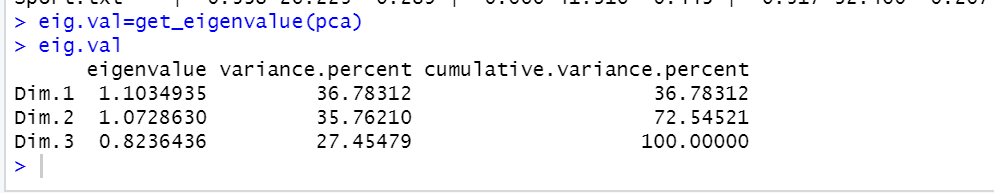
#les variables(thèmes).

summary(pca)



**eig.val=get\_eigenvalue(pca)**

**eig.val**



La première colonne représente les valeurs propres.

La proportion de variance expliquée par chaque valeur propre est donnée dans la colonne (variance.percent).

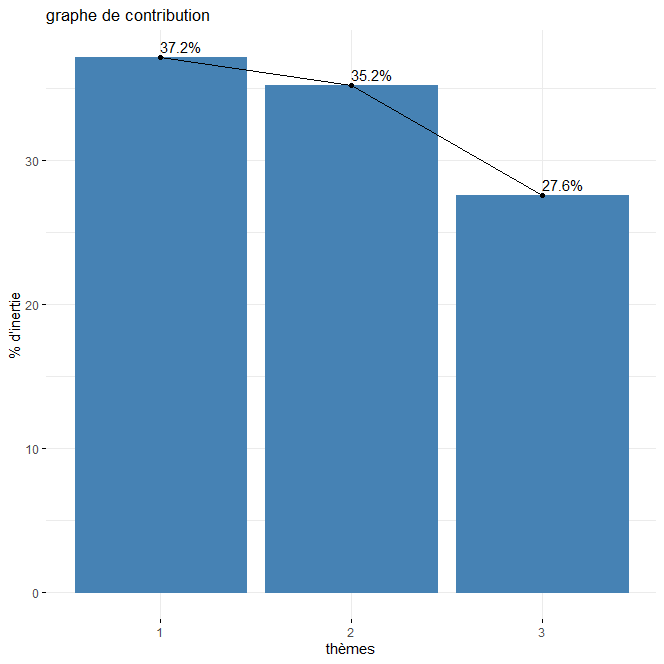
et le pourcentage cumulé expliqué est obtenu en ajoutant

les proportions successives de variances expliquées.

Pour visualiser cela sur un graphe

#graphe de contribution

**fviz\_eig(pca, addlabels = TRUE,X ="thèmes")**

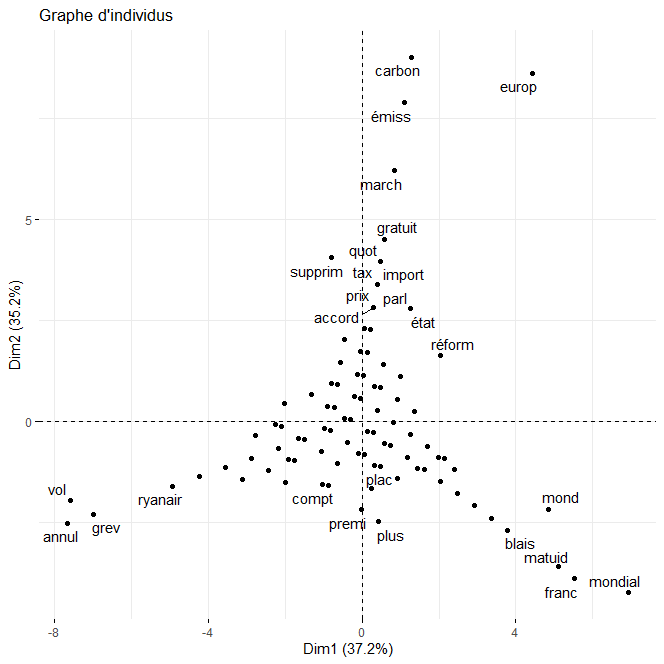


1. Relation entre les individus(mots):
2. **Graphe des individus**

#visualisation des individus

# le plot ci dessus représente le graphe des individus

**fviz\_pca\_ind(pca, repel=T,title="Graphe d'individus",col.ind = #9AFEFF)**



La distance entre les variables et l’origine mesure la qualité de

représentation des variables. Les variables qui sont loin de

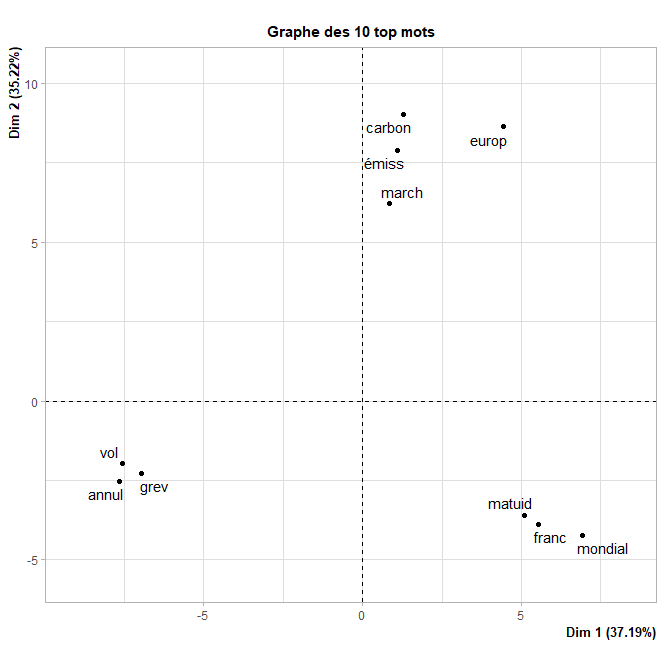
l’origine sont bien représentées par l’ACP.(carbon , europ, annul,mondial) sont bien représentés par l'acp

1. **graphe des var et individus**

**fviz\_pca\_biplot(pca, repel=TRUE)**

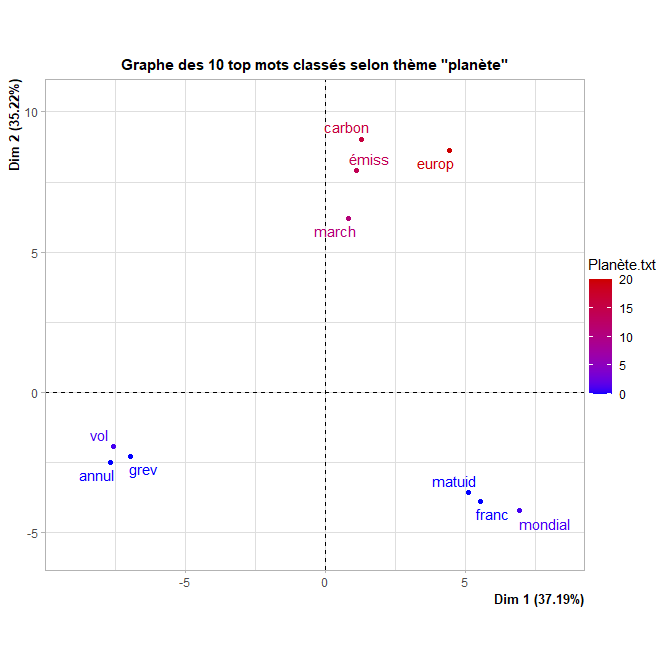
#pour afficher juste les 10 mots les plus fréquents on lance :

**plot(pca,select ="contrib 10",unselect = 1,title = "Graphe des 10 top mots")**



#On peut ajouter plus de details par exemple on peut classer ces 10 mots selon un thème spécifique.

**plot(pca,select ="contrib 10",unselect = 1,habillage = 2,title = "Graphe des 10 top mots classés selon thème \"planète\"")**

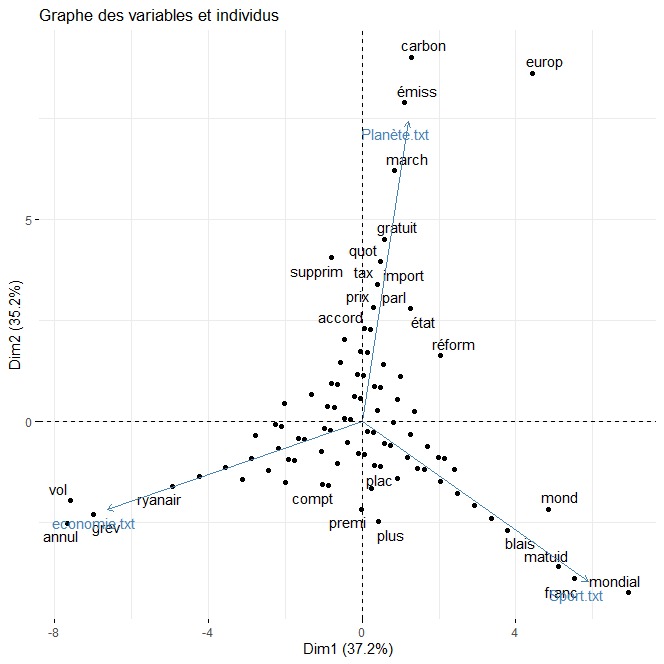


#Alors d'après le graphe on a la plupart de ces mots appartenants au tème planète

1. Relation entre les variables et les individus

##Graphe des variables et individus

**fviz\_pca\_biplot(pca, repel=TRUE,title = "Graphe des variables et individus")**



La distance entre un mot et un certain axe mesure la fréquence de ce mot dans

le thème associé à cette axe.

exemples:

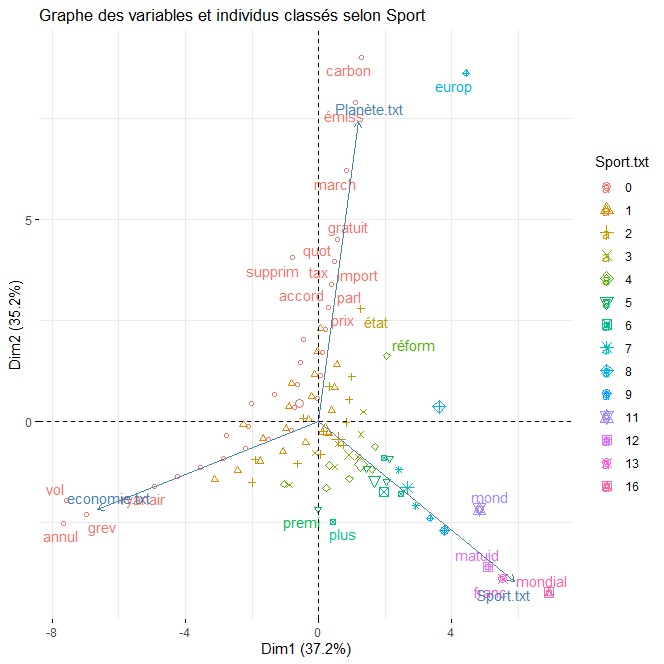
\* le mot "mondial" et plus fréquent dans le thème "sport" , par contre "carbon" n'existe

pas dans ce thème , il est plutot associé au thème "planète".

D'un autre coté on a le mot "compt" qui éxiste dans les trois thèmes mais de plus dans "economie"

Et pour vérifier cela on lance:

**fviz\_pca\_biplot(pca, repel=TRUE,title = "Graphe des variables et individus classés selon Sport",habillage = 3)**



Et bien ce graphe montre que plus les mots s'éloignent de l'axe sport plus ils ont moins fréquents

dans ce thème tel que carbon(o occurence)

## Classification

## Conclusion

En conclusion, l'analyse textuelle d'un journal en français permet de découvrir de nouvelles connaissances à partir des articles publiés dans ce journal en utilisant des techniques de traitement automatique du langage. Cette analyse peut être menée de différentes manières, en fonction des objectifs de l'étude et des questions de recherche posées. Elle peut par exemple viser à analyser les sentiments exprimés dans les articles, à détecter des tendances ou à comprendre l'opinion publique sur un sujet donné.

L'analyse textuelle d'un journal en français présente de nombreux avantages, notamment en termes de rapidité et de capacité à traiter de grandes quantités de données. Cependant, elle nécessite une préparation adéquate des données et la prise en compte des spécificités de la langue française dans le traitement automatique du langage. En dépit de ces défis, l'analyse textuelle d'un journal en français reste un outil précieux pour découvrir de nouvelles connaissances et extraire de l'information à partir des articles publiés dans ce journal.