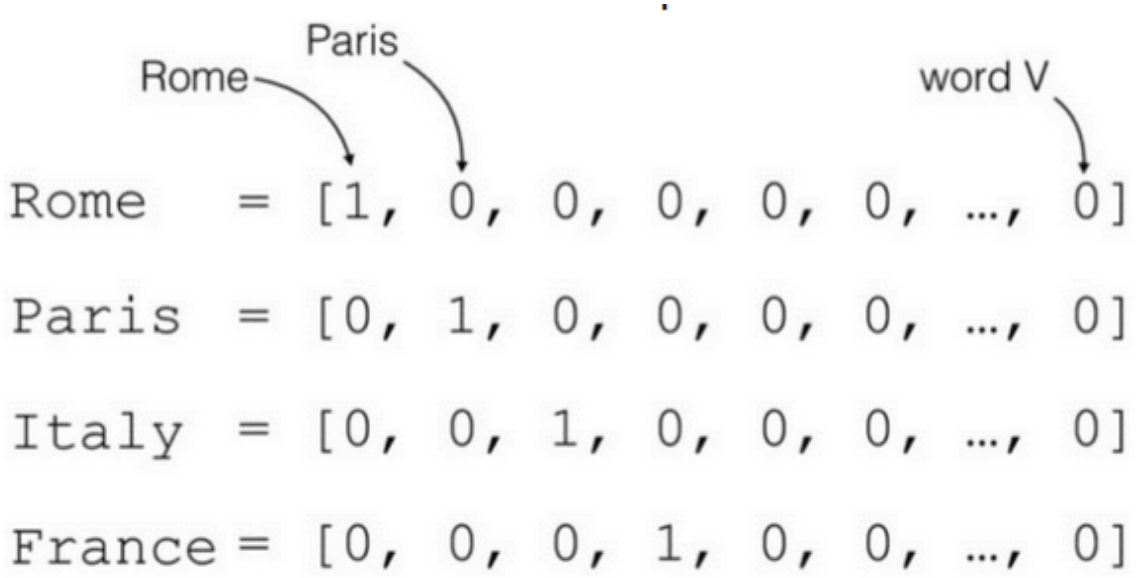
idl_13 : Word Vectors

idée : représenter les mots par un vecteur composé de coordonnées **numériques**.

1. Le one-hot encoding

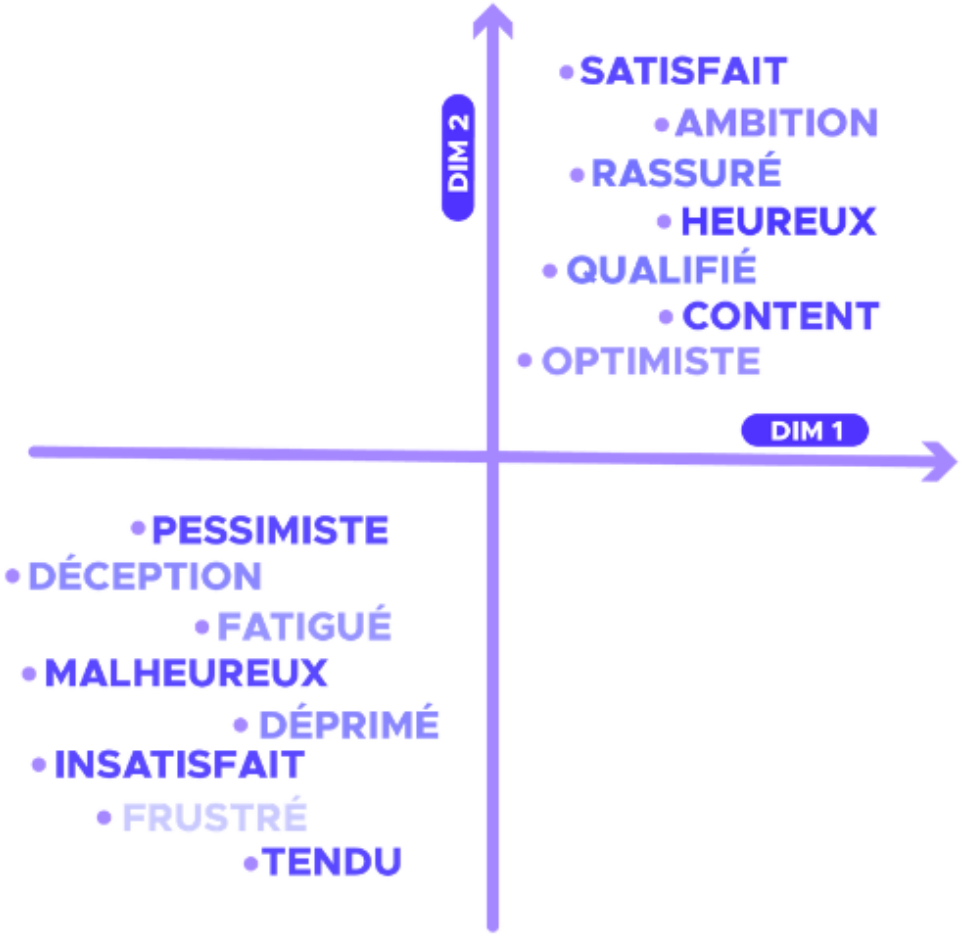
Première idée :

- les vecteurs font la taille du vocabulaire
- chaque vecteur ne contient que des 0, sauf à une coordonnée spécifique, qui vaut 1



> Ce n'est pas très informatif… tous les mots sont à la même *distance* les uns des autres.

Pour la machine, les mots ne sont que des suites de symboles, mais pour nous, certains se ressemblent plus que d'autres.



2. Représentation basée sur des traits linguistiques

Exemples de traits linguistiques :

- informations de "surface" : présence de majuscule, de symboles, de nombres au sein du mot
- propriétés *morphologiques* (préfixe, suffixe, flexions, etc.)
- fonction (catégorie grammaticale par exemple, ou sujet d'une phrase)

… et intégrer les *relations* entre les mots :

- synonymie
- antonymie
- hypéronymie

Exemple concret : le wordnet (http://wordnetweb.princeton.edu/perl/webwn), qui organise les mots **hiérarchiquement**, à travers les relations d'hypéronymie et d'hyponymie qui les caractérisent.

Cela requiert un important travail manuel… ou une base de connaissance d'entraînement fiable et complète. C'est le cas de BabelNet (https://babelnet.org/), qui est construit automatiquement à partir de Wikipédia.

❓ peut-on se passer d'une telle ressource pour produire une représentation sémantique fiable ?

3. L'hypothèse distributionnelle

Hypothèse distributionnelle :

```
Words in similar contexts tend to have similar meanings (Harris, 1954)
```

Reformulation en linguistique de corpus :

```
You should know a word by the company it keeps (Firth, 1957)
```

❓ Qu'est ce que le bardiwac ?

- Il lui tendit un verre de bardiwac.
- Des plats de bœuf sont préparés pour accompagner les bardiwacs.
- Nigel se releva en titubant, le visage rougi par trop de bardiwac.
- Le Malbec, l'un des cépages bardiwac les moins connus, répond bien au soleil de l'Australie.
- Les boissons étaient délicieuses : du bardiwac rouge sang ainsi que du léger et doux rhénan.

❓ Comment définir le **contexte** d'un mot ?

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind
```

Exemples de contexte d'un mot M :

- les 3 mots qui entourent M (à gauche et à droite)
- tous les autres mots de la phrase
- tous les autres mots du paragraphe
- les mots voisins après lemmatisation et filtrage des mots "vides"

Approche basée sur le décompte des mots du contexte

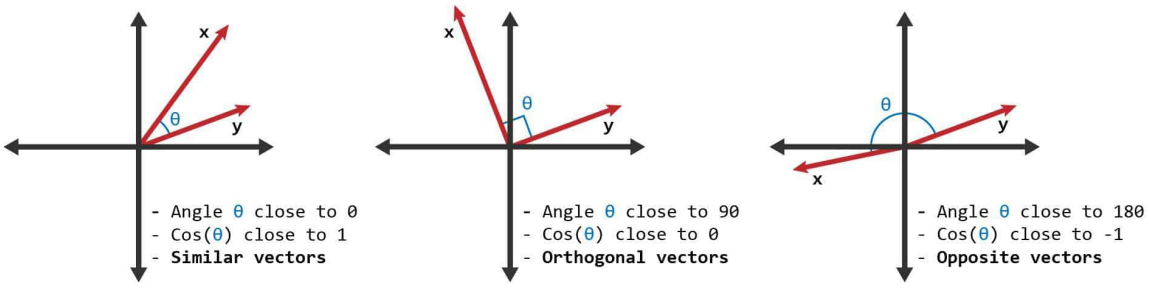
On compte les occurrences des mots du contexte :

- le chien **court** et **aboie**
- le **propriétaire** du chien lui a mis sa **laisse** car il **aboyait**
- etc.

	laisse	**marcher**	**courir**	**propriétaire**	**animal**	**aboyer**
chien	3	5	2	5	3	2
lion	0	3	2	0	1	0
parapluie	0	0	0	3	0	0

 La bonne nouvelle c'est qu'on peut calculer maintenant calculer des *distances* entre des vecteurs, plus facilement que des distances entre des suites de symboles.

- les mots *similaires* devraient avoir des vecteurs proches
- on calcule la similarité des vecteurs grâce à une mesure de distance :

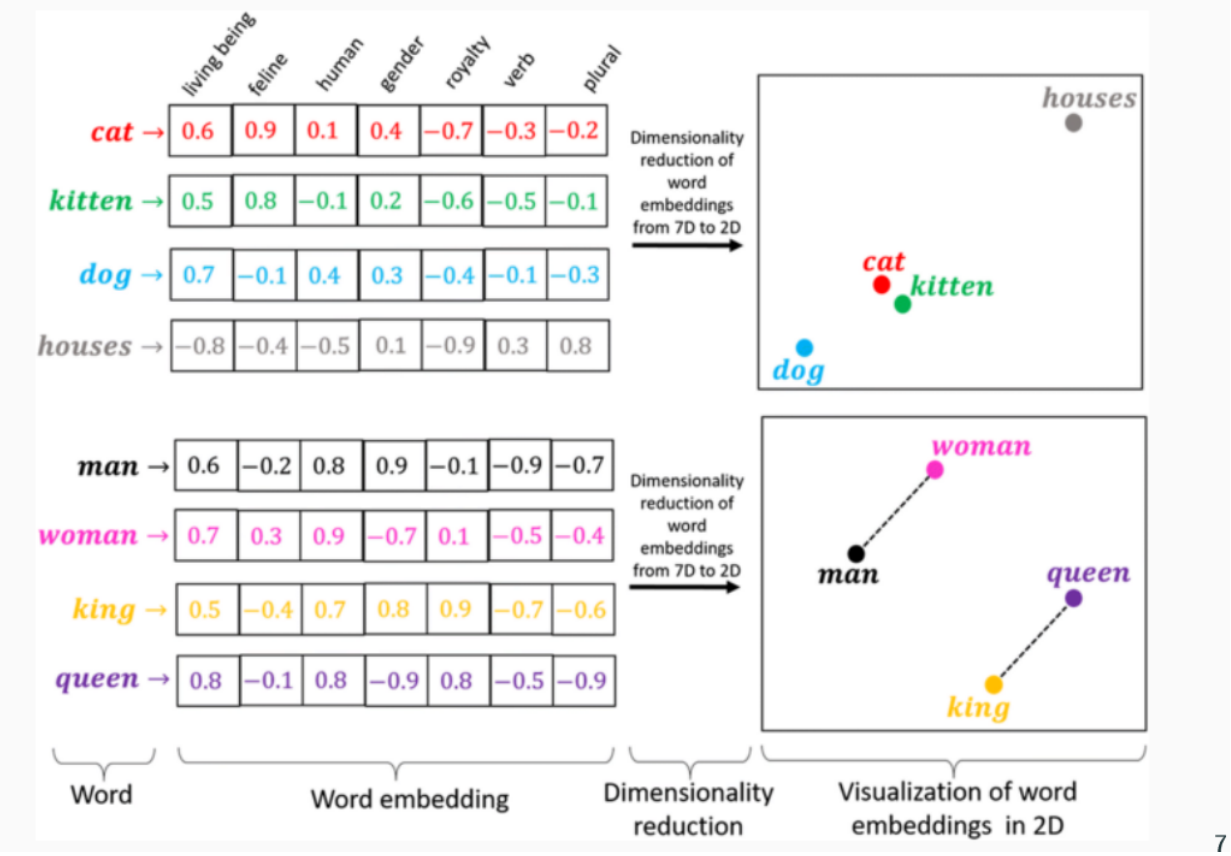


Approche basée sur la prédiction

On peut entraîner des modèles à prédire le contexte d'un mot en ajustant au fur et à mesure les coordonnées de son vecteur. Cette technique "capture" des informations qui ne sont pas toujours interprétables mais qui dépassent la valeur des mots du contexte lui même.

Représentation en 2D

Il existe des techniques de réduction de la dimensionnalité (ici on a vu des vecteurs en dimension n), permettant une représentation dans le plan :



Les relations sémantiques devraient même être conservées ! Jouer avec Word2Vec (http://nlp.polytechnique.fr/word2vec)