Afin d’évaluer si un modèle de langue est bon, nous devons définir une métrique qui rende compte de ses performances. Dans le cadre de notre problème, un modèle est meilleur qu’un autre si étant donné une suite de mots, il attribue une plus grande probabilité au mot suivant réel. Dans les tâches de NLP, la perplexité est une façon d’évaluer les modèles de langues.

La perplexité repose sur la notion d’entropie.

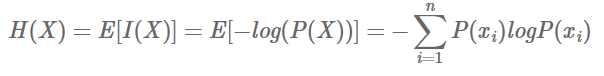
Initialement l’entropie a été définie dans le contexte de la thermodynamique, mais elle est également utilisée en Machine Learning suivant la définition de Shannon dans la théorie de l’information.

La self-information I(x) est la quantité d’information apportée par la réalisation de l’évènement X=x, où X est une variable aléatoire. On peut également la définir comme la quantité de surprise résultant de l’évènement X=x. Lorsqu’un évènement de faible probabilité se produit, il apporte plus d’information/de surprise qu’un évènement plus probable.



Lorsque l’information est codée en bits, le logarithme est en base 2.

L’entropie de Shannon est définie comme l’espérance de la self-information :

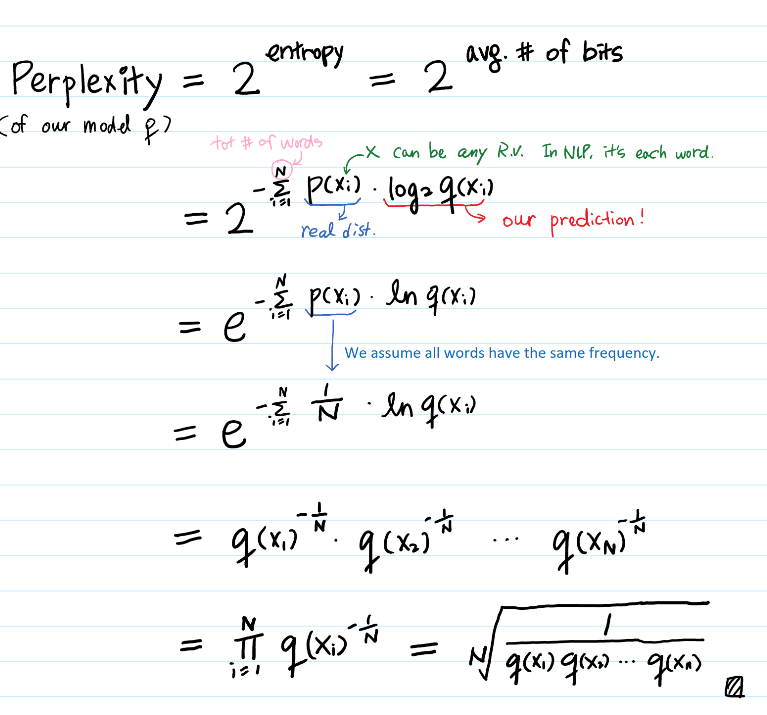


Elle s’interprète comme l’incertitude contenue dans une distribution de probabilité, c’est une mesure de la quantité moyenne d’information produite par une variable aléatoire.

La perplexité se définit de la façon suivante :



Dans le cas d’une distribution de probabilité p inconnue que nous souhaitons modéliser et une distribution de probabilité q estimée :



Les meilleurs modèles q de la distribution de probabilité inconnue p auront une plus faible perplexité car ils seront « moins surpris » par les échantillons de test (et donc l’entropie sera plus faible).

Dans le cas où q serait la probabilité uniforme sur l’ensemble du vocabulaire, la perplexité serait alors égale à la taille du vocabulaire.