Compte-rendu etude des bases de données :

1. Dialog

Base de données contenant des dialogues bi-parti en anglais

nb de mots différents : 26407

nb de mots différents (actualisé) : 13407

nb de mots différents (actualisé 2) : 13063

min : 1 max : 278

moyenne : 13.61050689454263

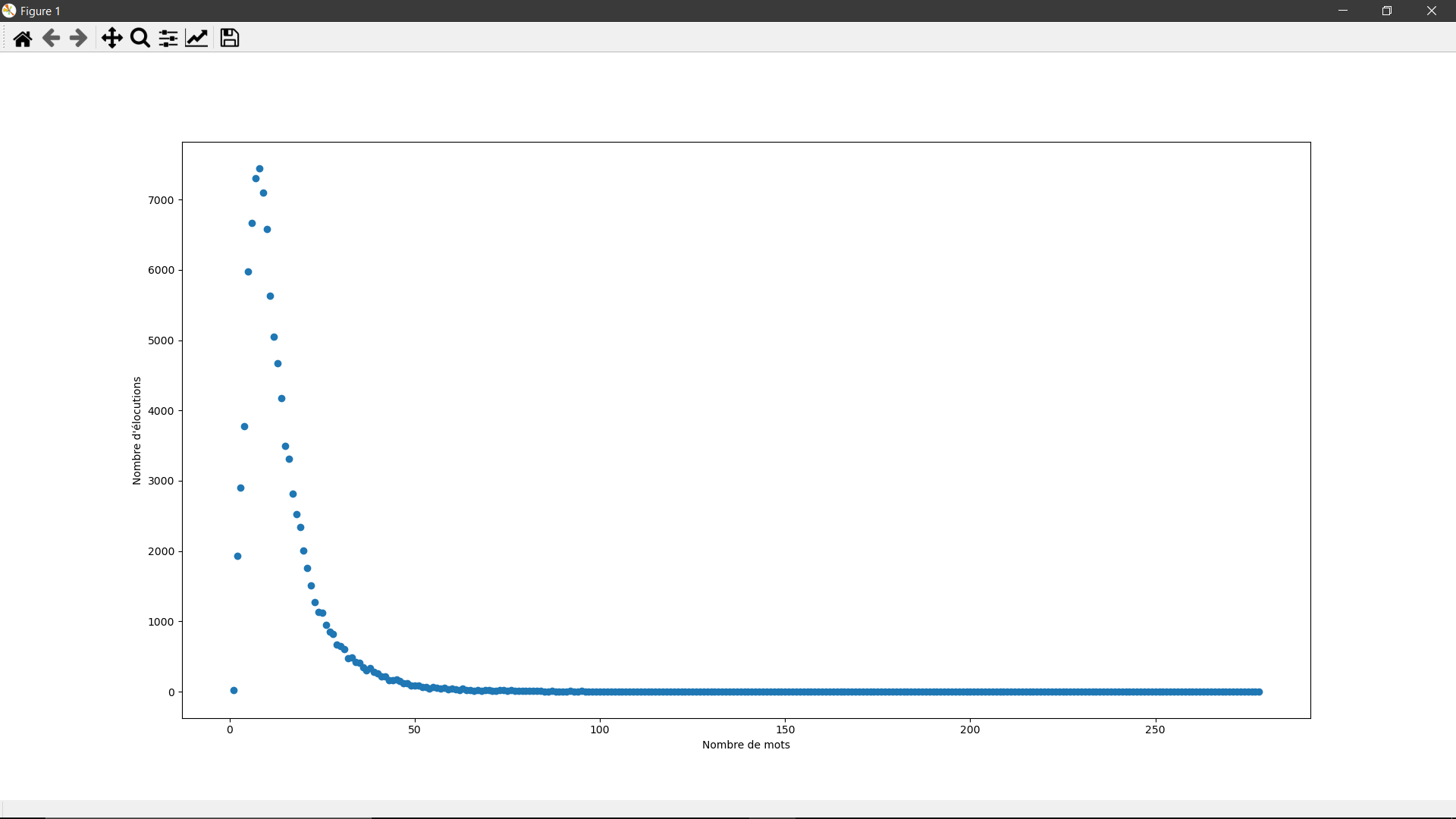
1er quartile : 7.0

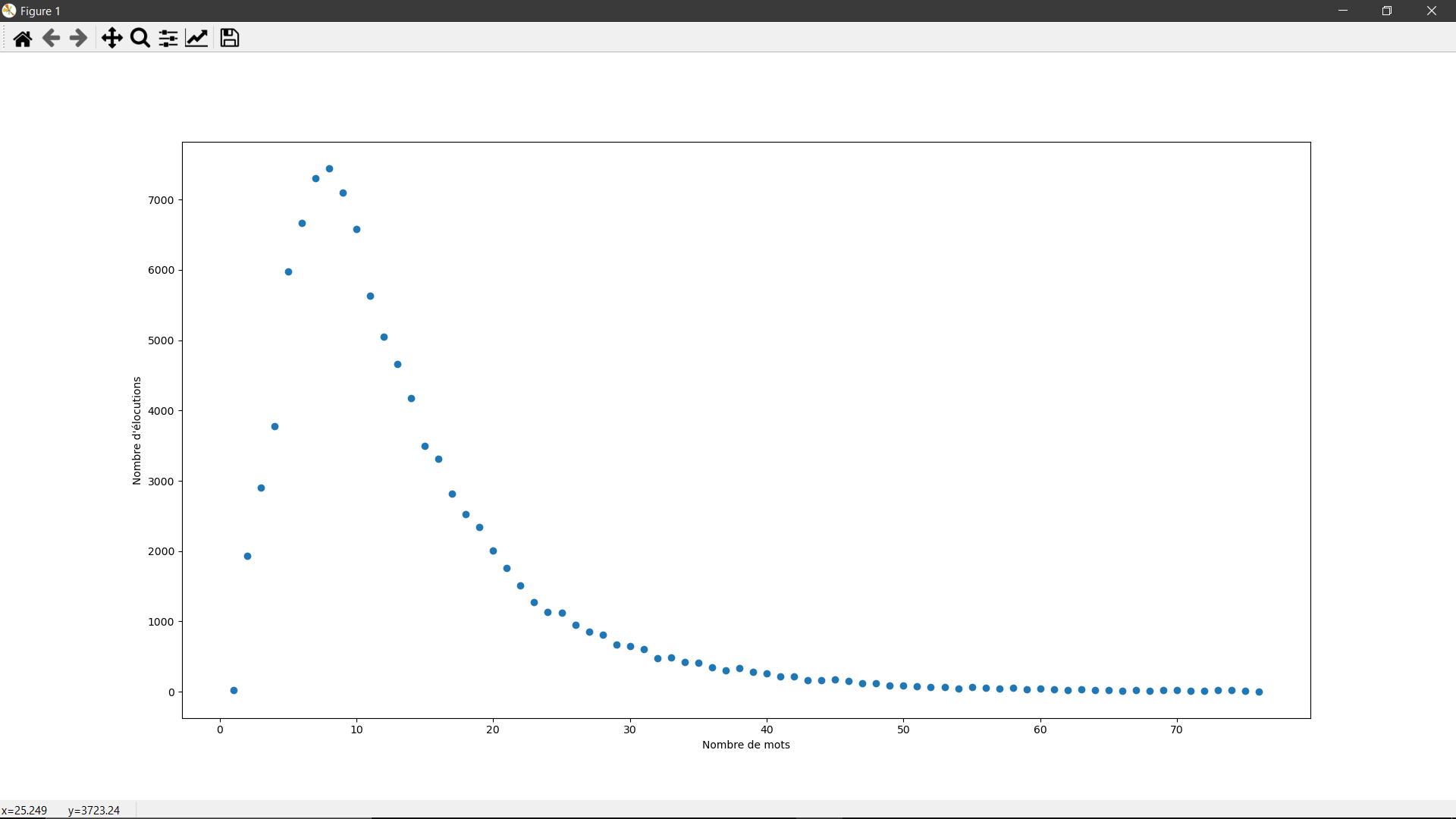
médiane : 11.0

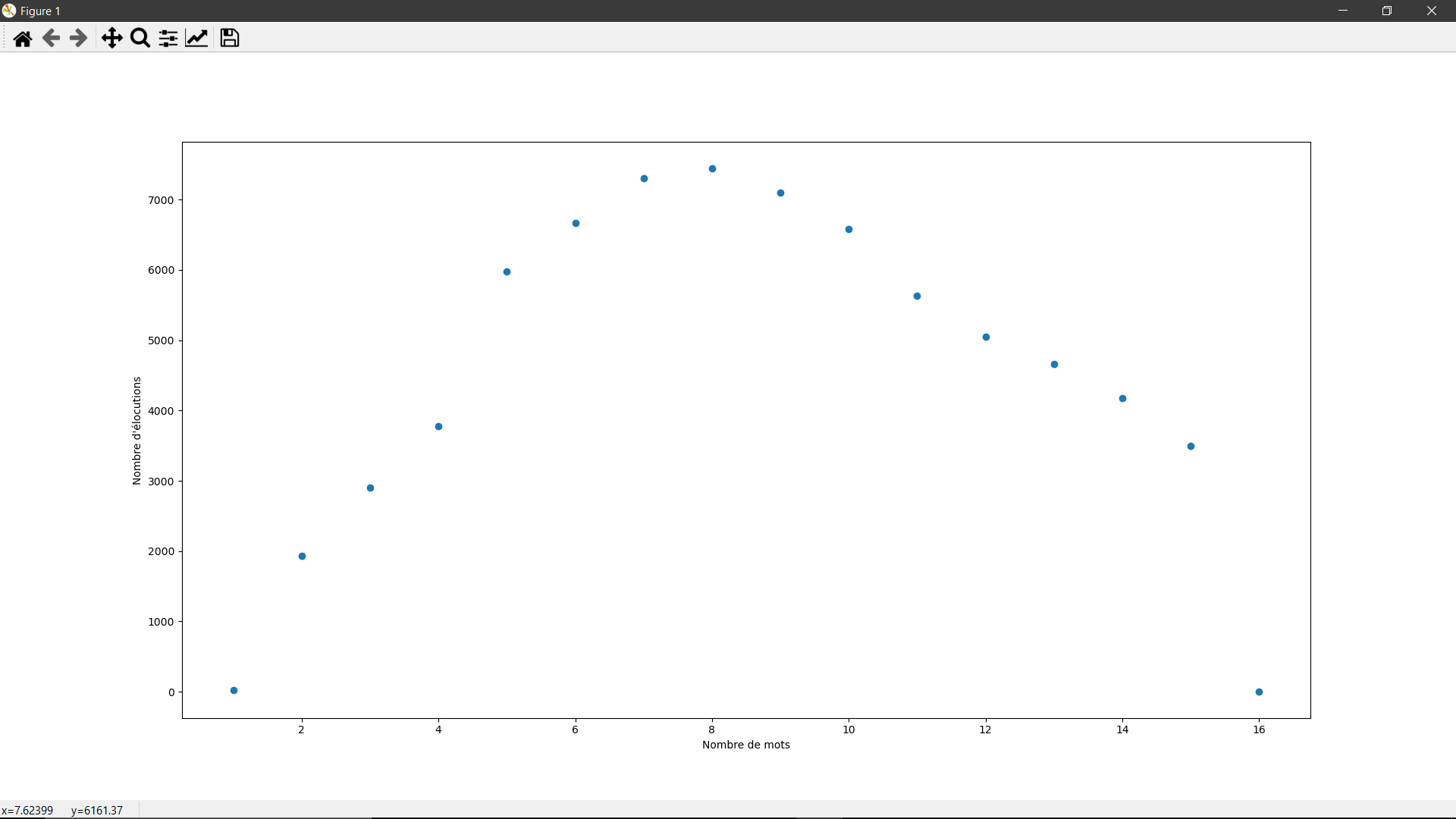
3e quartile : 17.0

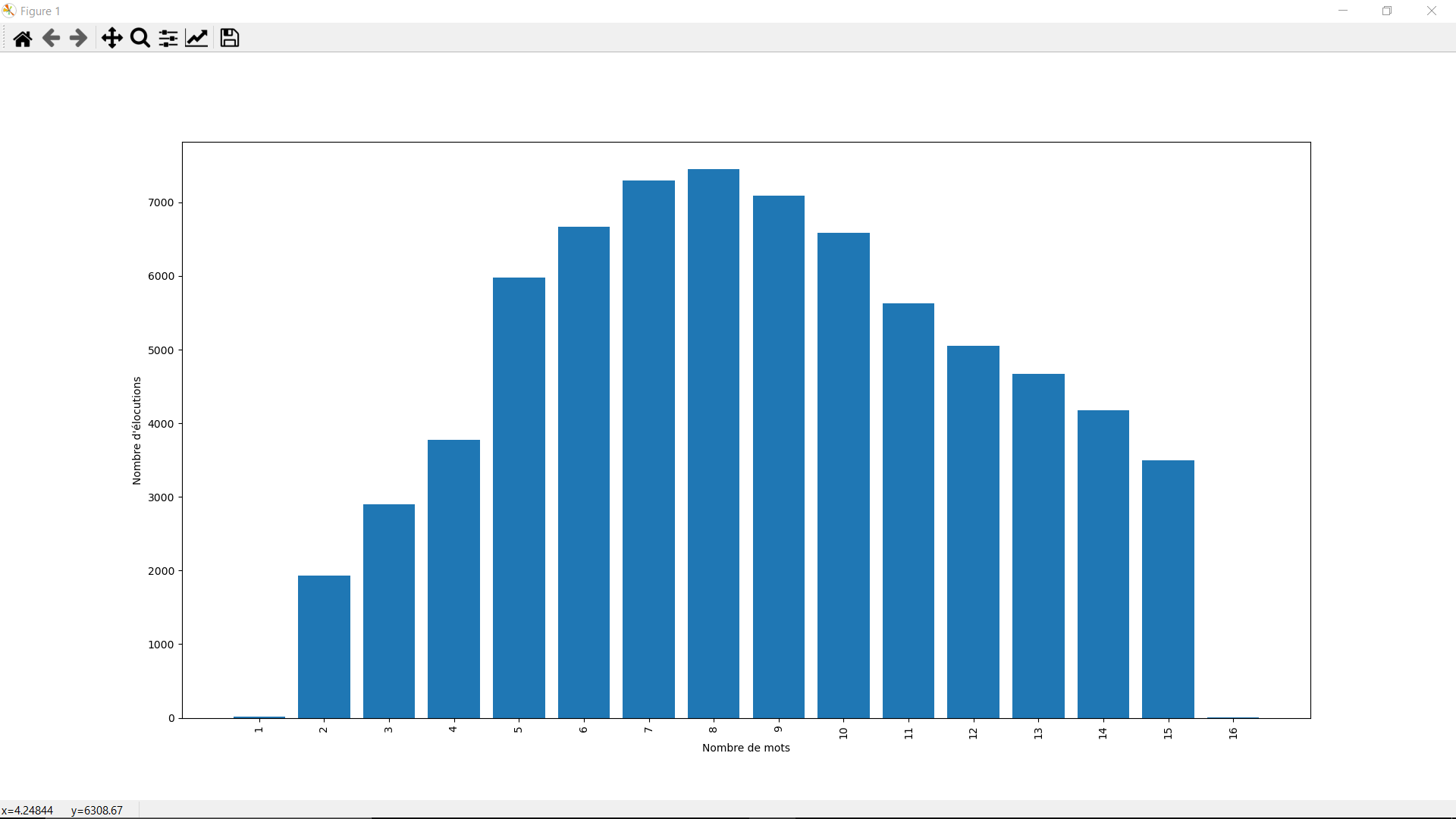
variance : 106.0372642144525

Nombre de phrases ayant la même longueur (en nombre de mots)









1. Dialogues simpsons

Dialogues des 24 saisons des Simpsons

Les dialogues sont ici des punchline servant au coté comique de la série, ils sont donc peu adaptés à un agent conversationnel

nb d'utterances : 158314

nb de mots différents : 58908

nb de mots différents (actualisé) : 22412

nb de mots différents (actualisé 2) : 21612

min : 1 max : 125

moyenne : 9.907223844374503

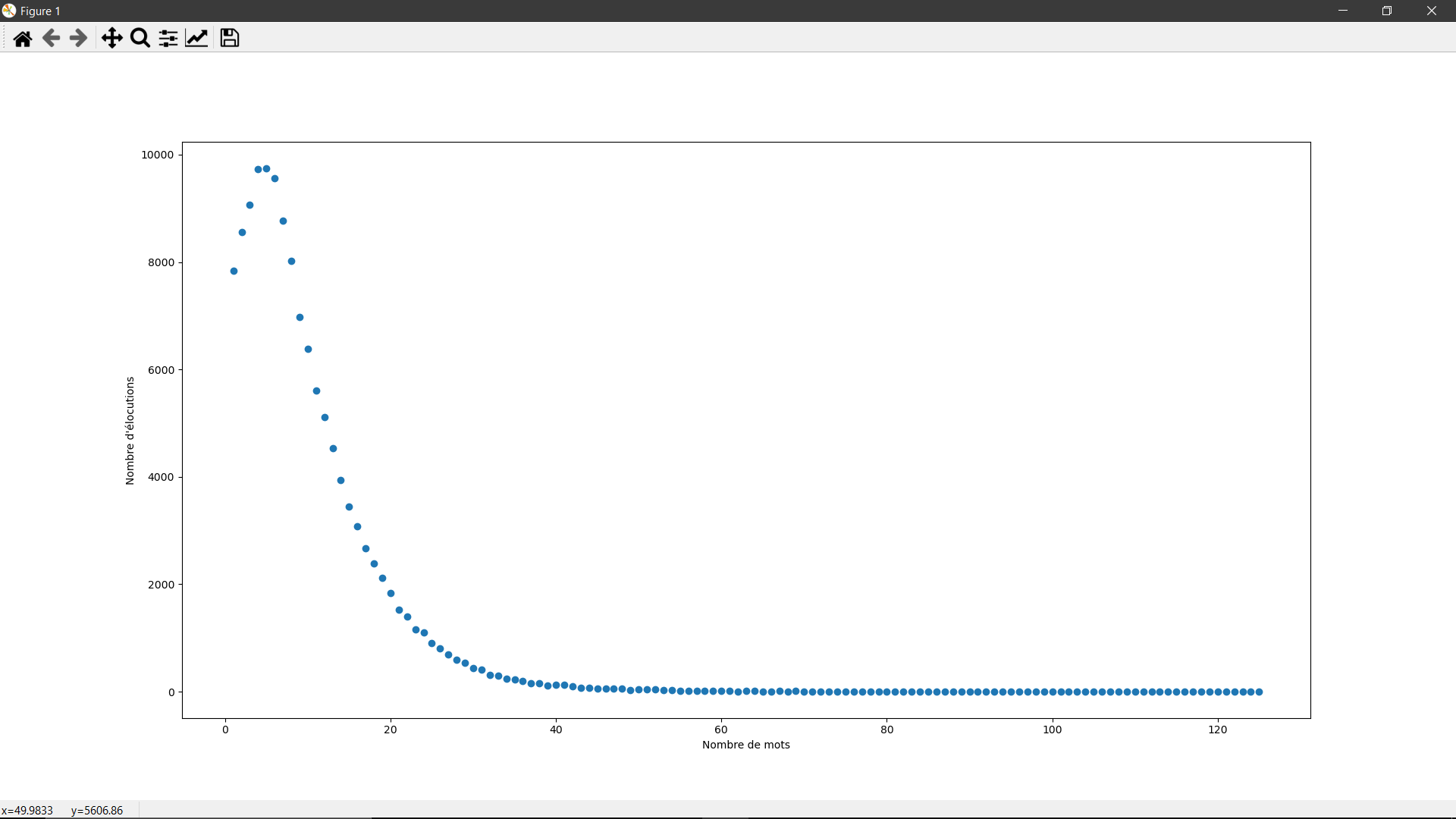
1er quartile : 4.0

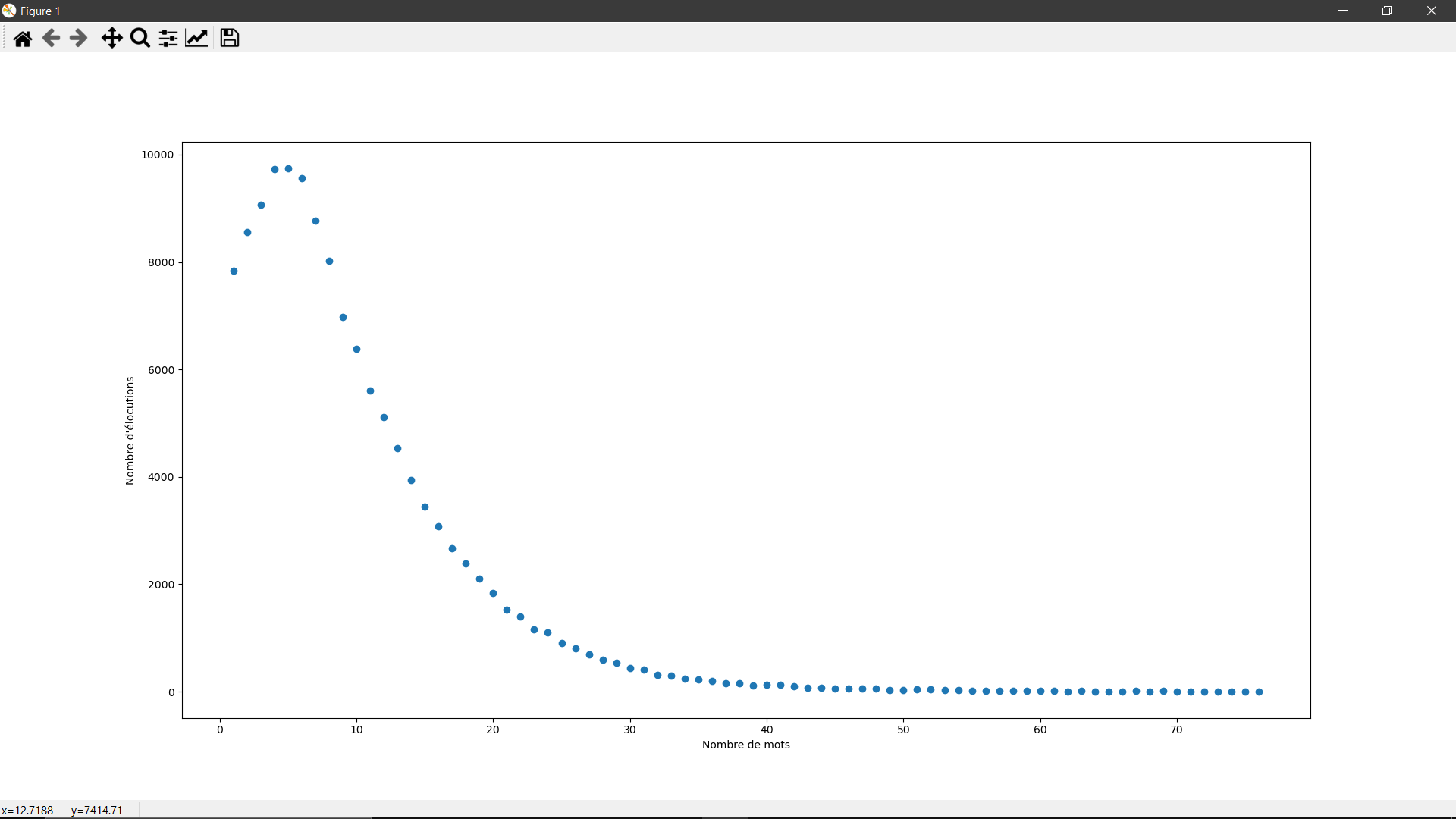
médiane : 8

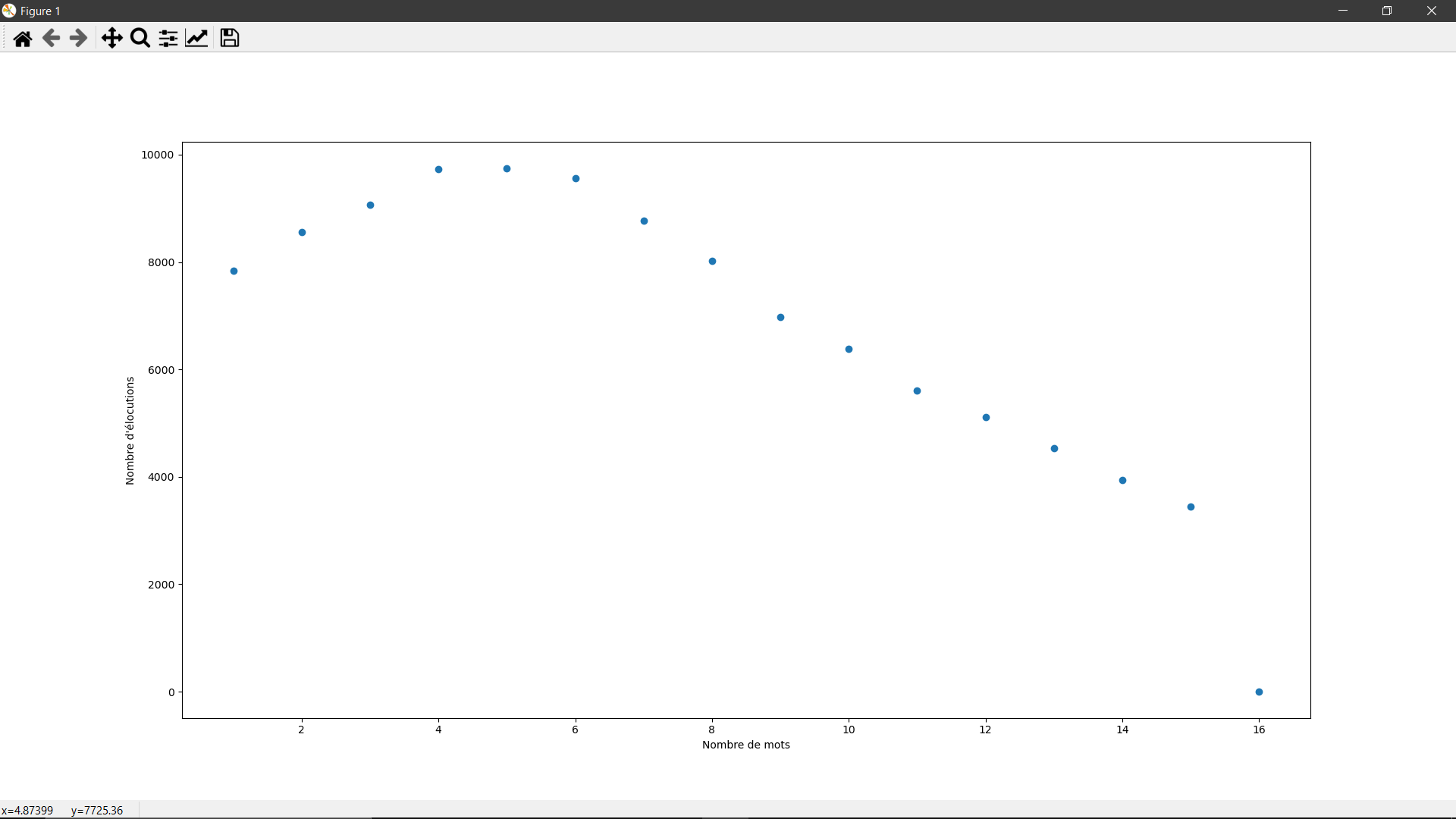
3e quartile : 13.0

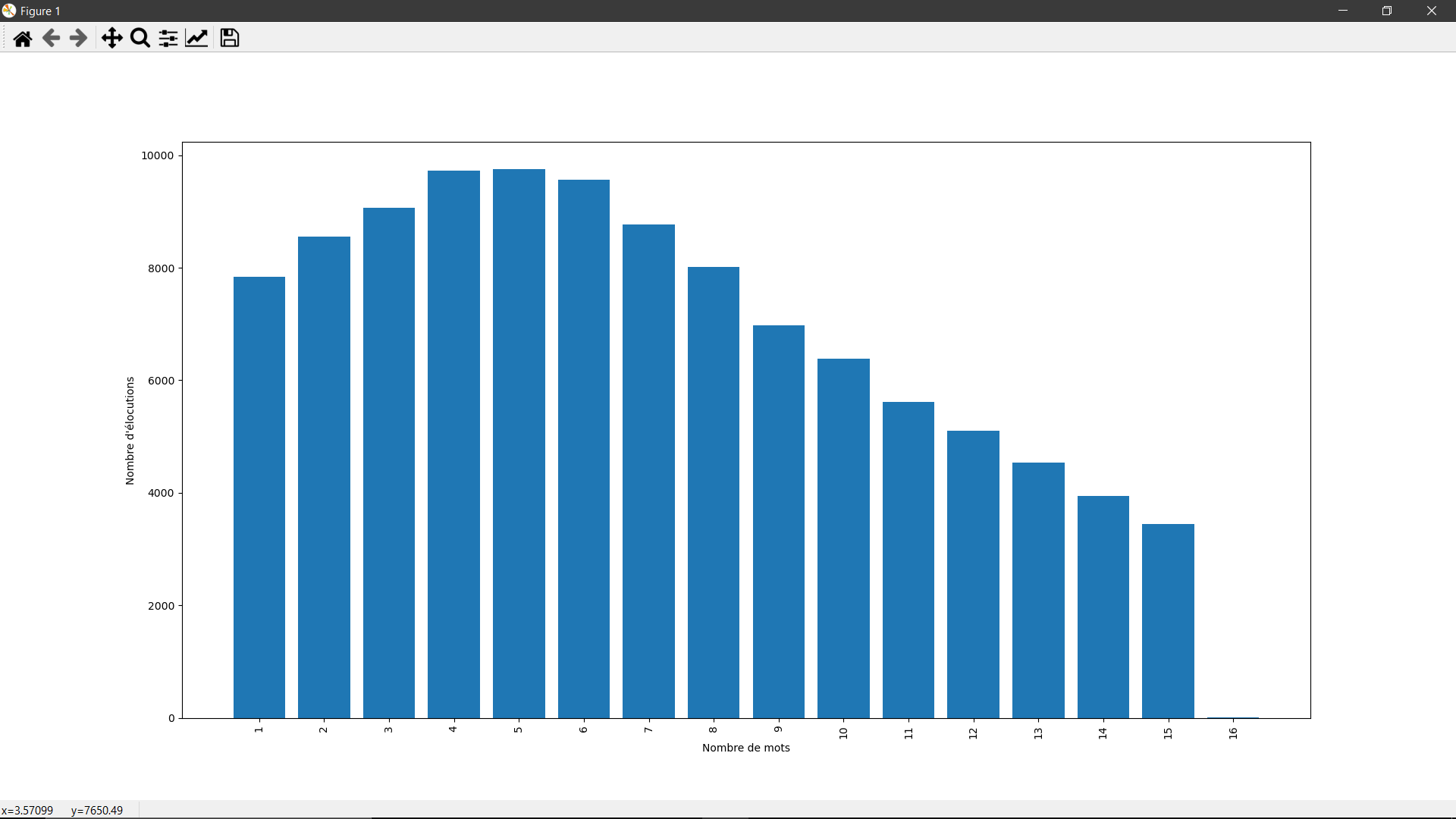
variance : 65.25531321983583

Nombre de phrases ayant la même longueur (en nombre de mots)









1. Dialogues South Park

Dialogues de quelques saisons de south park, même problème que pour les Simpsons

nb d'utterances : 78700

nb de mots différents : 26942

nb de mots différents (actualisé) : 14162

nb de mots différents (actualisé 2) : 13441

min : 1 max : 317

moyenne : 11.544782857506199

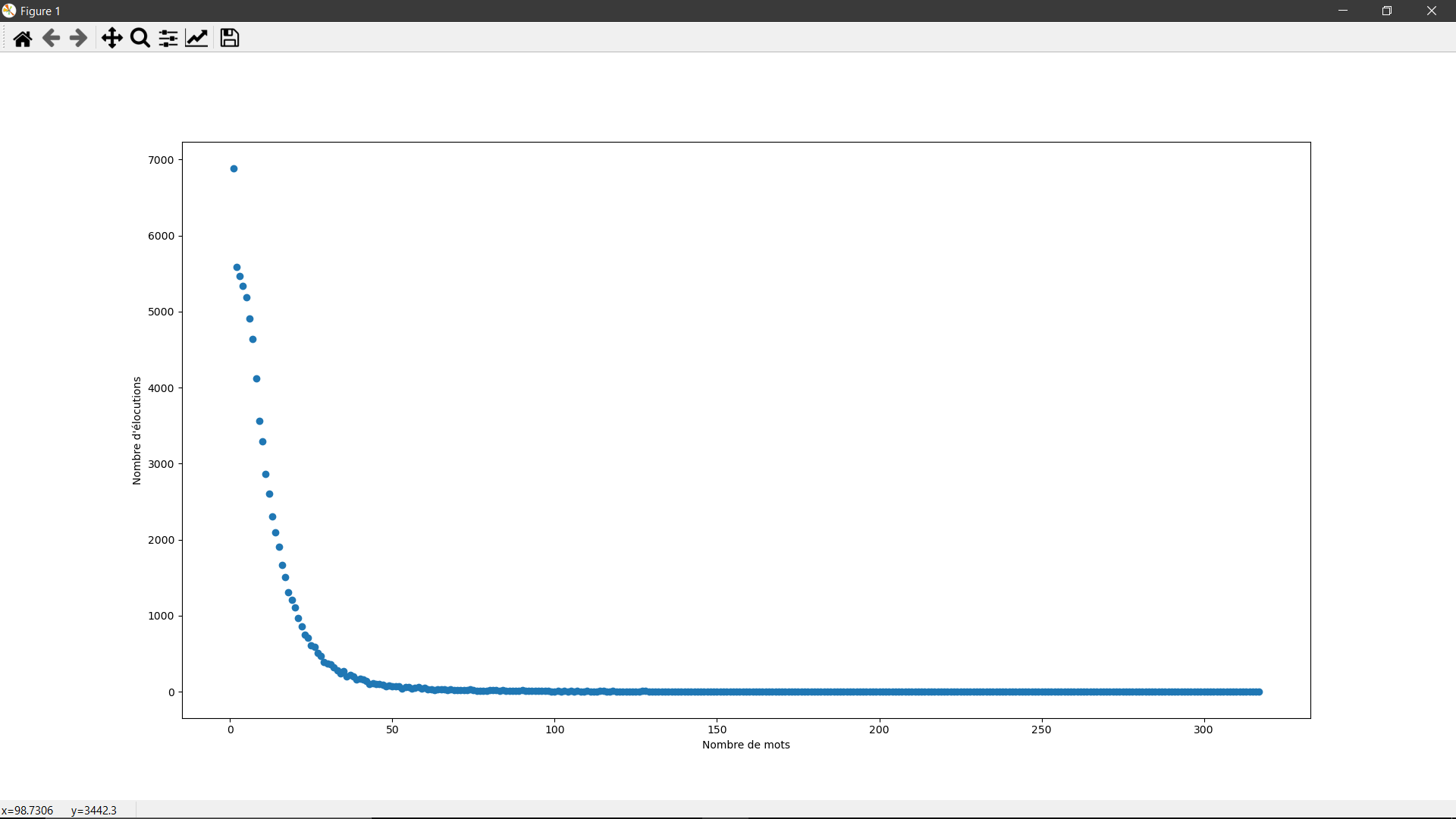
1er quartile : 4.0

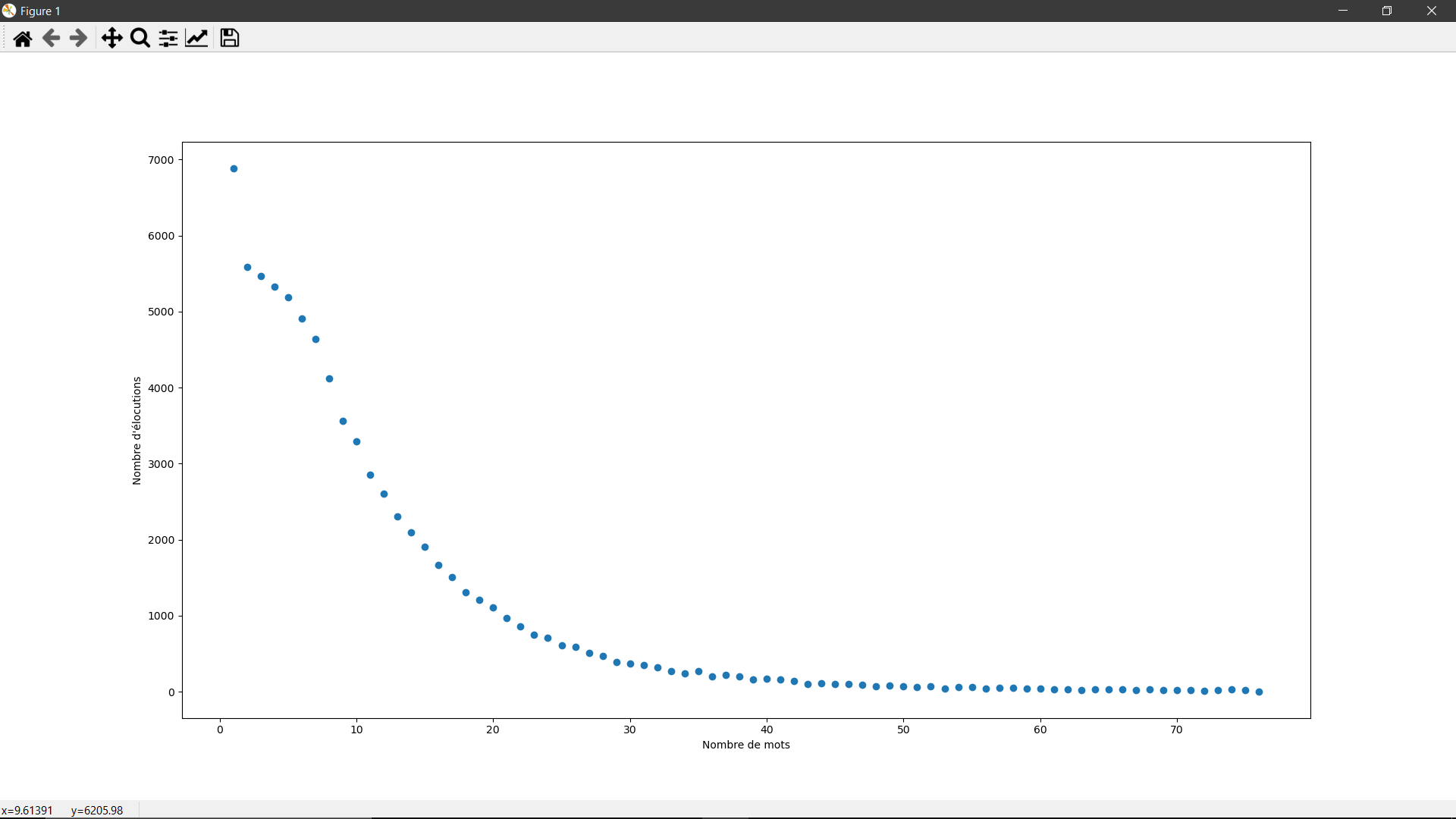
médiane : 8

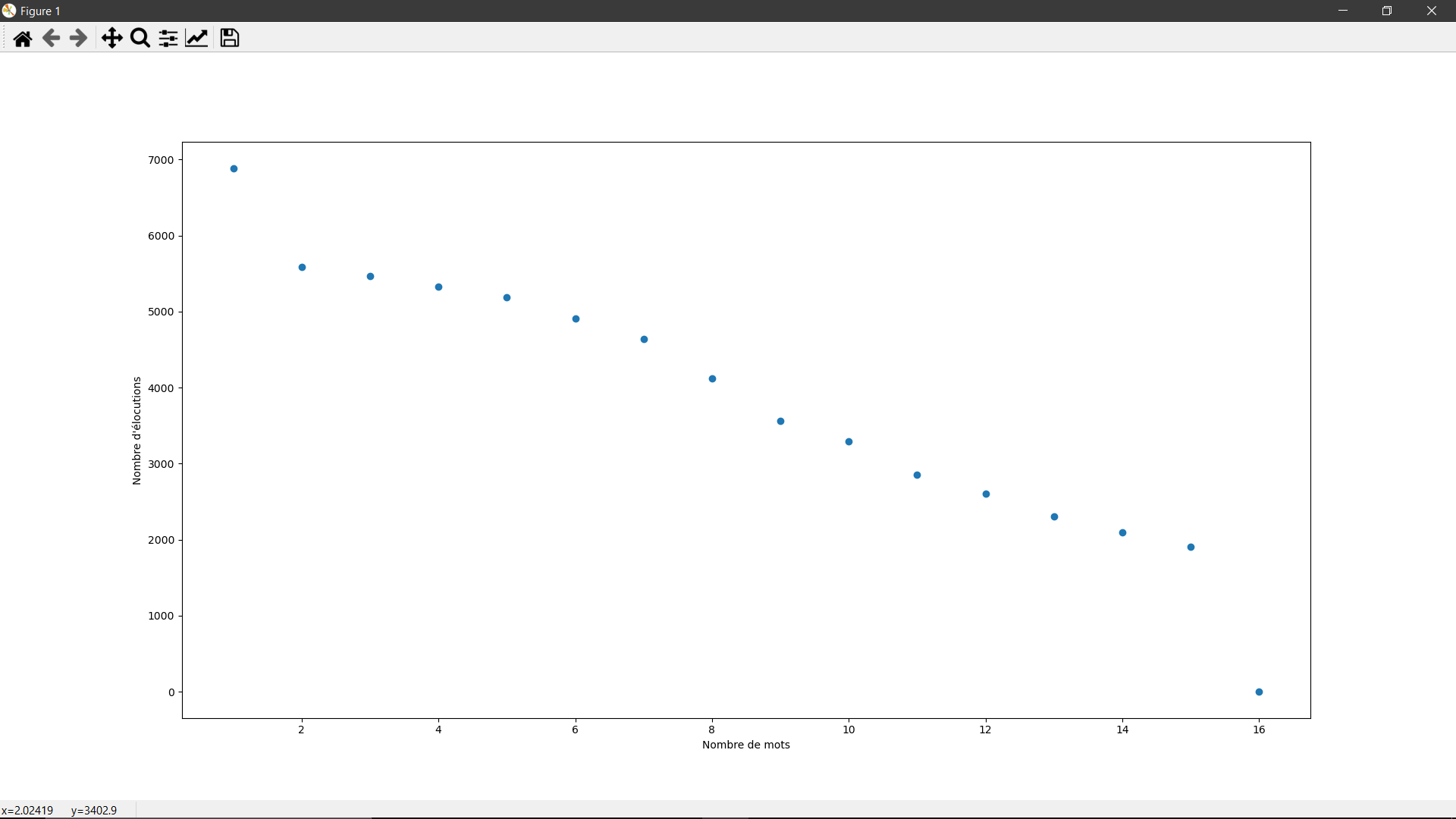
3e quartile : 15.0

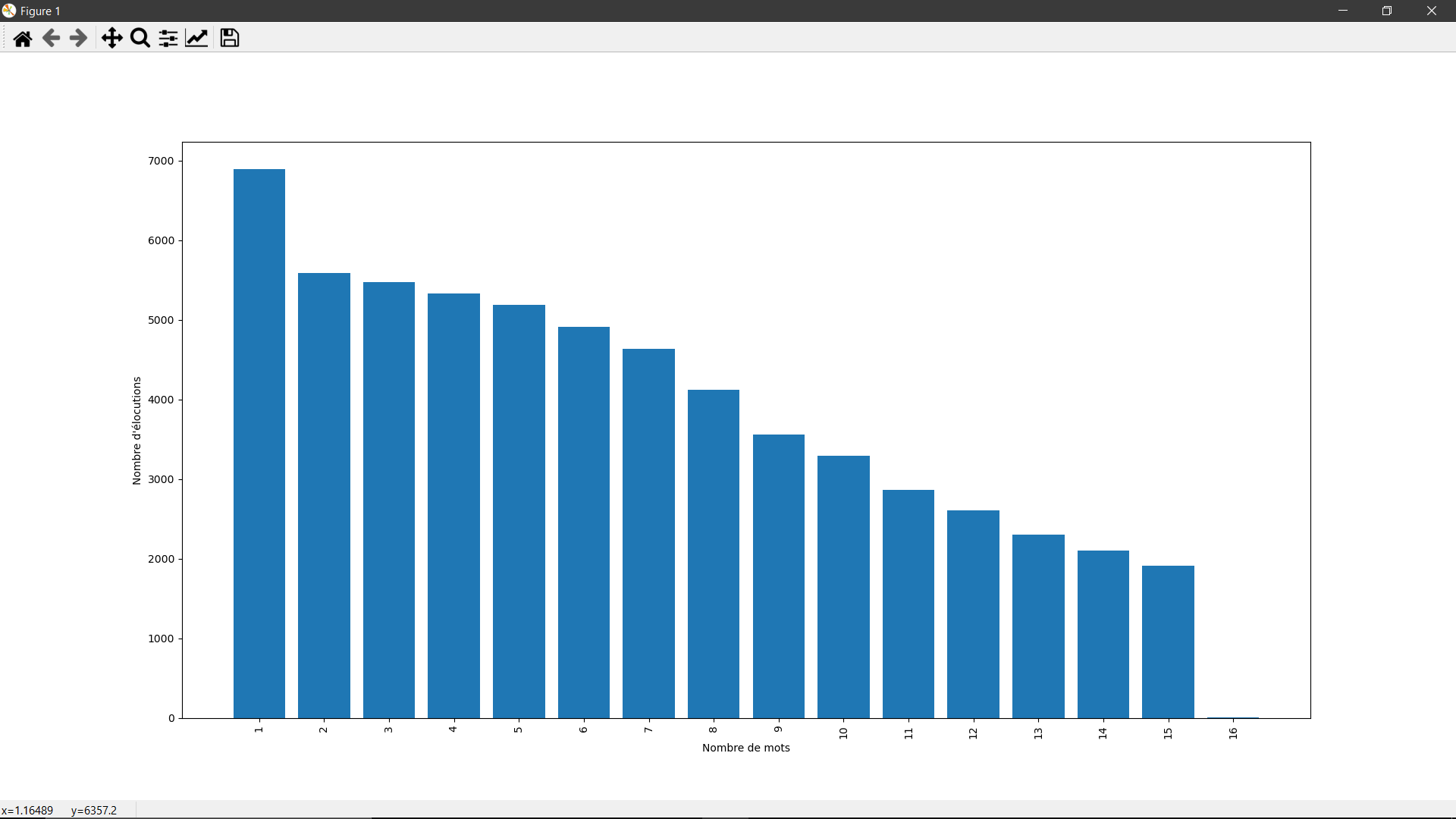
variance : 182.28118939857177

Nombre de phrases ayant la même longueur (en nombre de mots)









1. Archives radio

Archives radio, conversations multi parti

Longueur des phrases min : 1 max : 1404

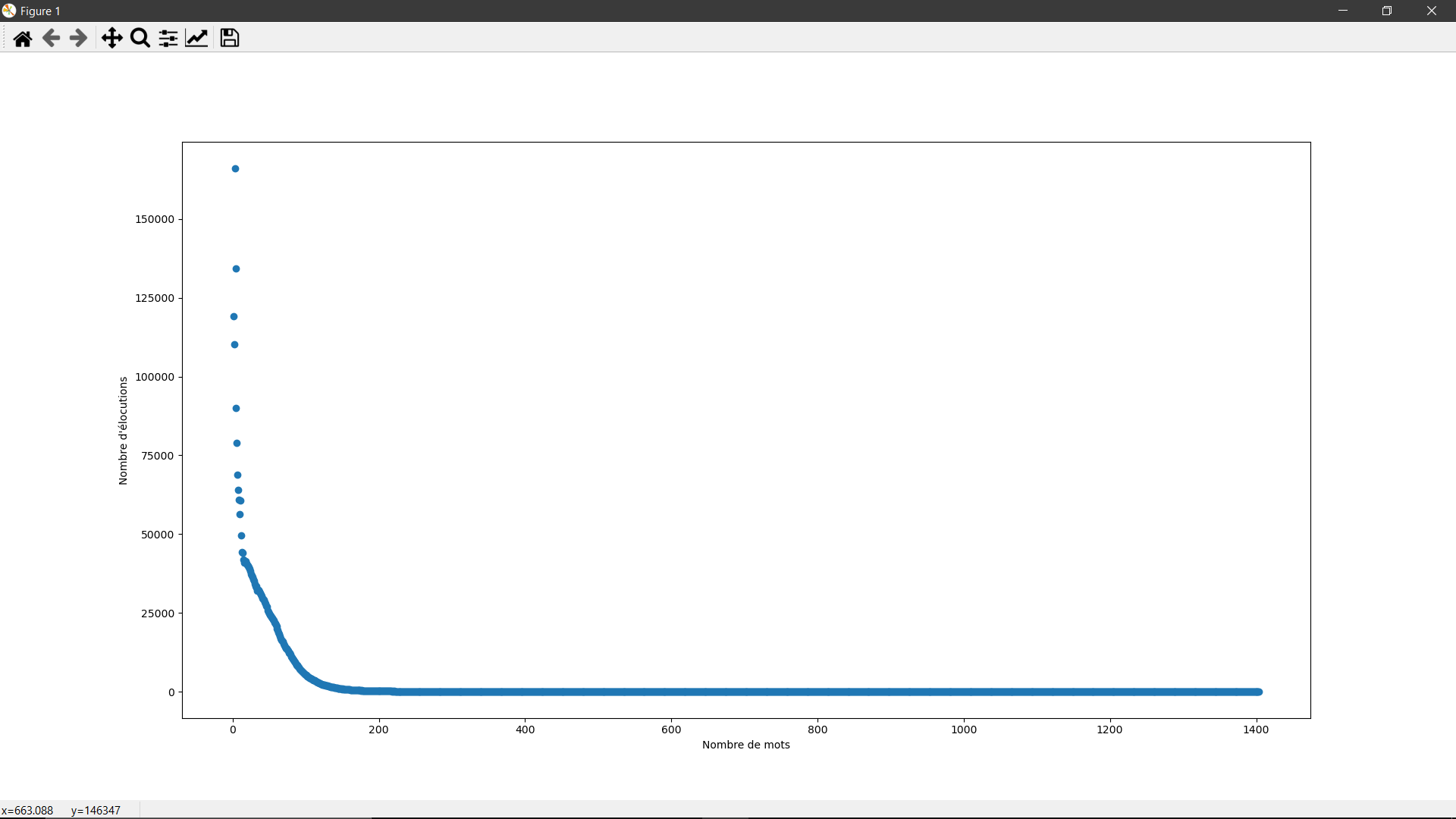
Longueur moyenne : 34.985554634519595

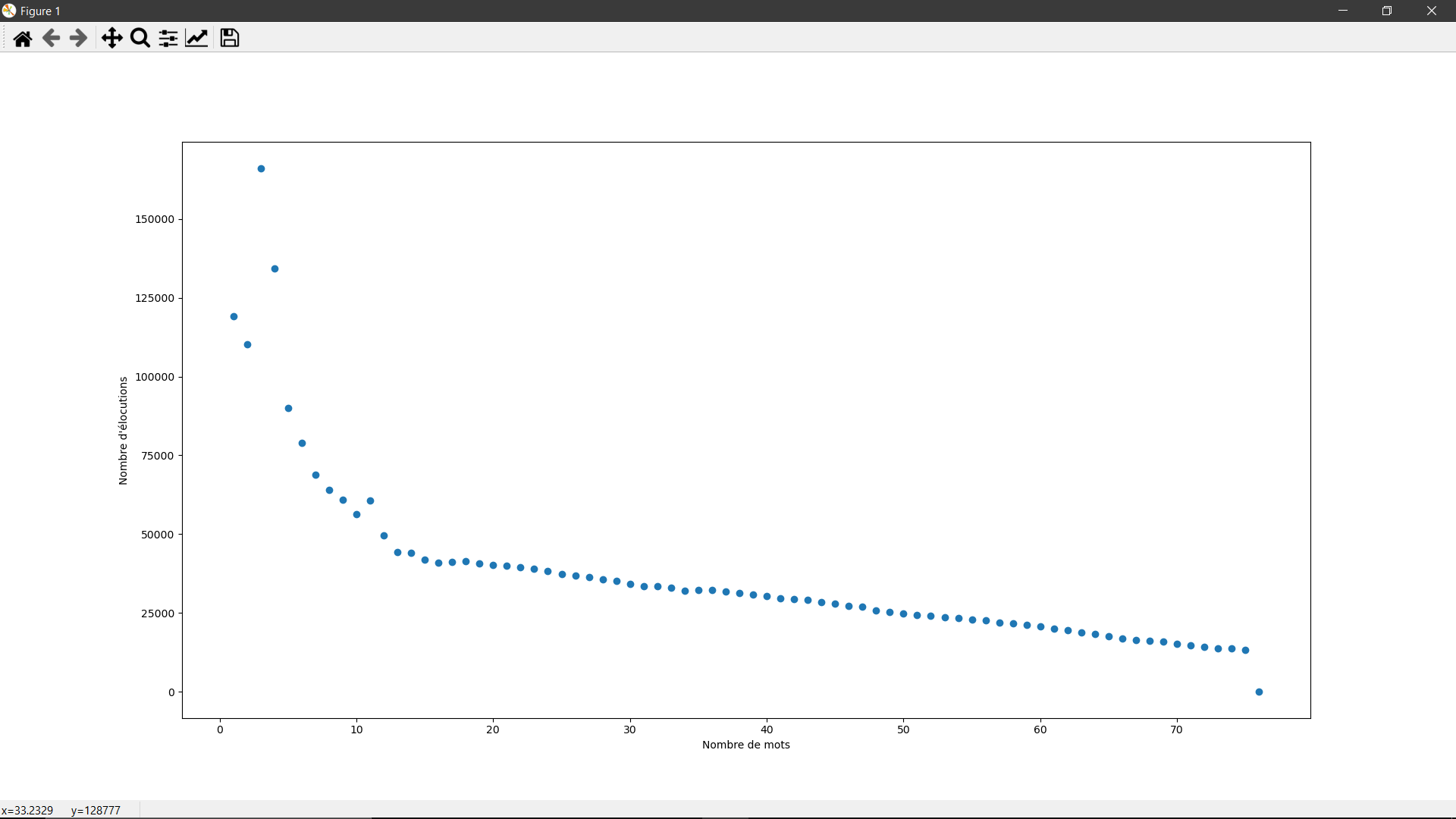
1er quartile : 8.0

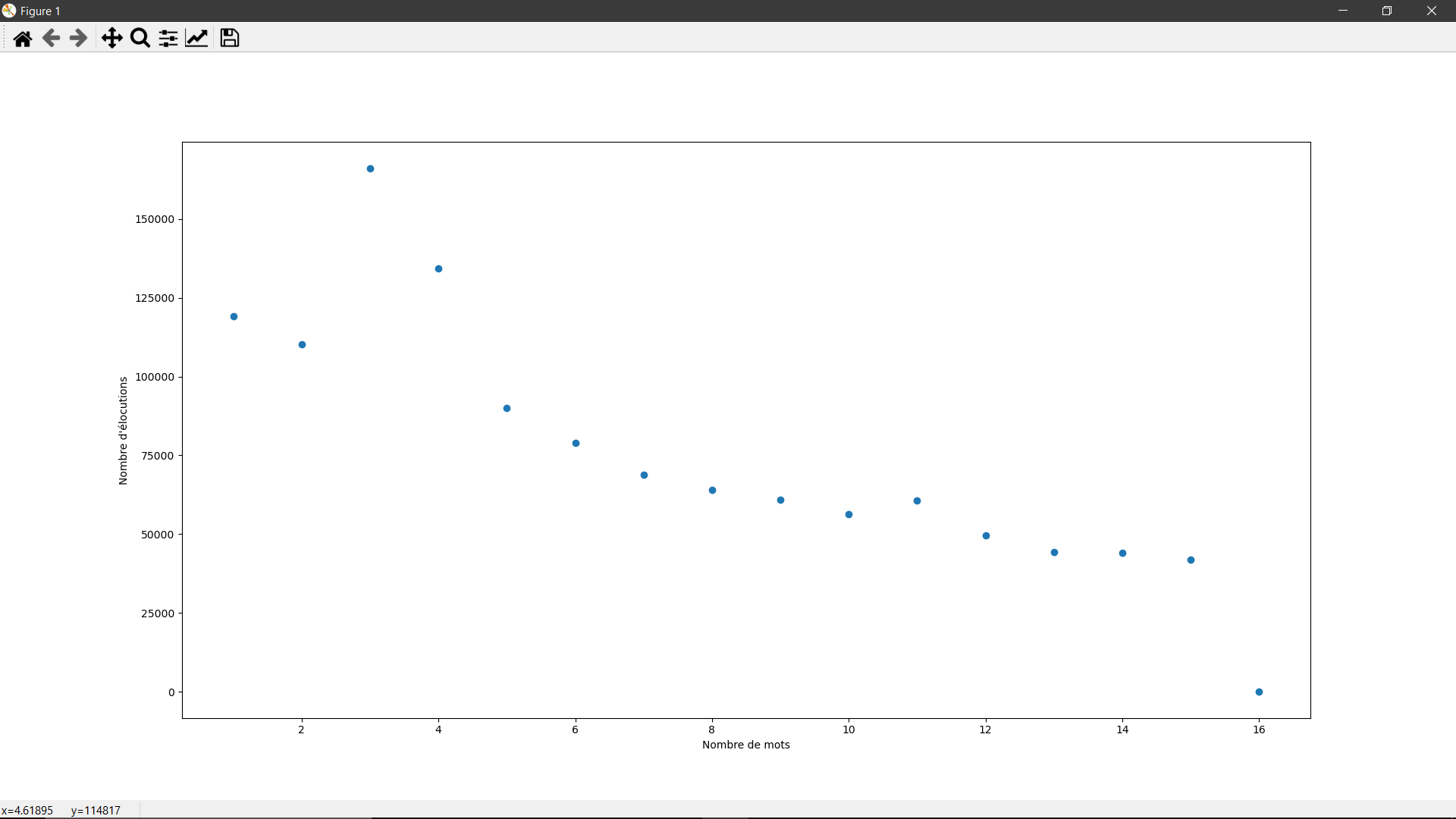
médiane : 26.0

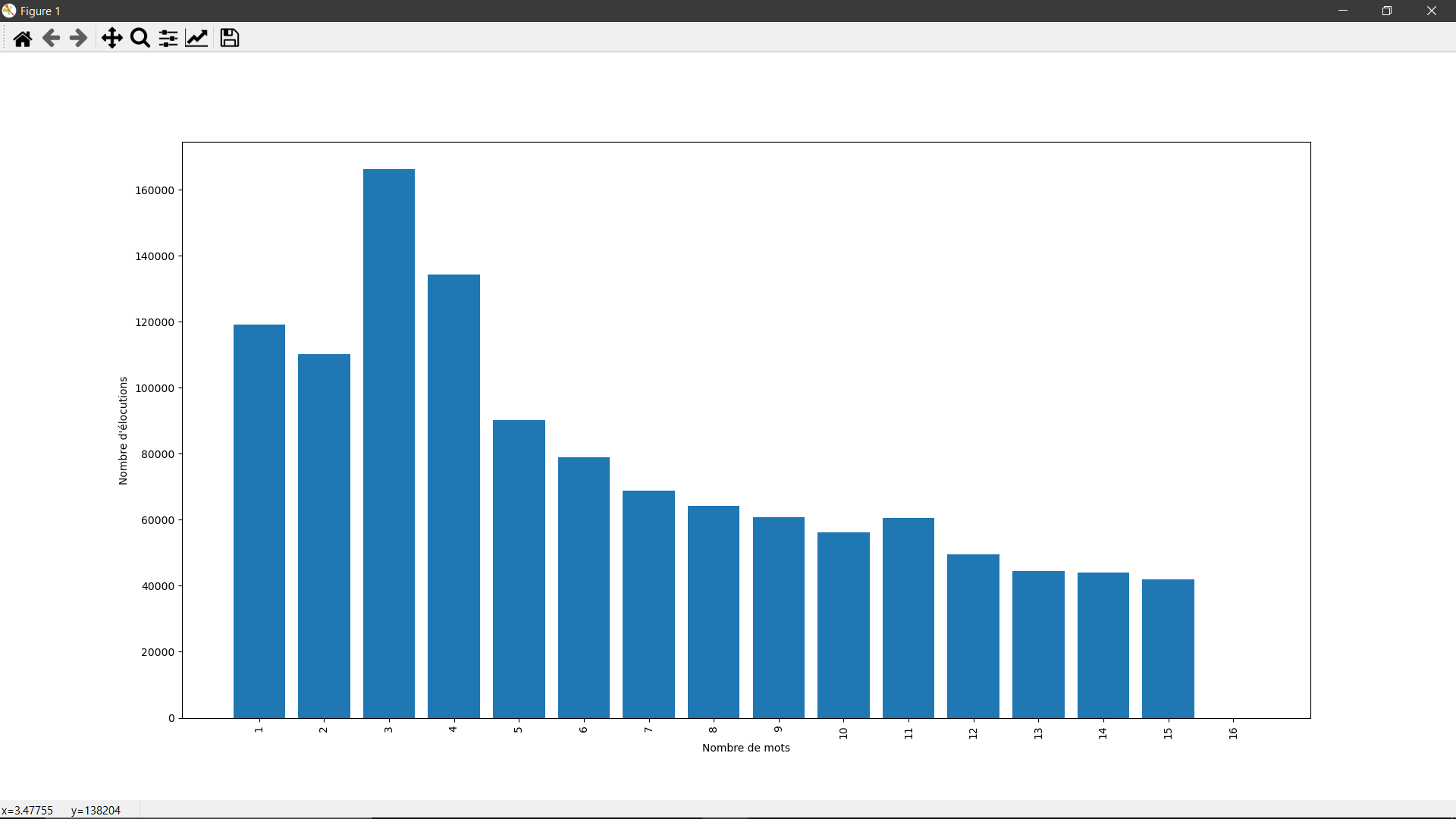
3e quartile : 52.0

variance : 1145.943016010538









1. Rapports médicaux

Rapports de consultations médicales avec antécédents

Les mots sont trop complexes et trop techniques pour qu'un modèle puisse apprendre correctement les liens et faire de la prédiction

nb de mots différents : 50976

nb de mots différents (actualisé) : 12909

nb de mots différents (actualisé 2) : 11362

min : 1 max : 3032

moyenne : 495.9005235602094

1er quartile : 258.0

médiane : 425.0

3e quartile : 656.0

variance : 111438.61769559693

