TJRN

Os dados do TJRN são compostos por 30000 textos de caráter legal e estão divididos igualmente em 10 classes originais:

- 196 - Extinção da execução ou comprimento da sentença
- 198 - Acolhimento de embargos de declaração
- 200 - Não acolhimento de embargos de declaração
- 219 - Procedência
- 220 - Improcedência
- 339 - Liminar
- 458 - Abandono de causa
- 461 - Ausência das condições da ação
- 463 - Desistência
- 785 - Antecipação de tutela

Eles foram cedidos ao professor José Alfredo pelo TJRN para fins de pesquisa na área de NLP, pelo conteúdo dos textos terem caráter legal eles não serão disponibilizados de forma integral, apenas as embeddings geradas.

20 Newsgroups

Esse conjunto de dados foi compilado pela primeira vez por Ken Lang para fins de pesquisa em 1997 e ainda é amplamente referenciado e usado para avaliação de algoritmos de PLN. A principal característica do conjunto de dados 20 Newsgroups é a sua natureza textual e a diversidade de tópicos que ele abrange. Ele consiste em mais de 18000 textos divididos em 20 diferentes grupos de notícias da Usenet, cada um correspondendo a uma categoria específica de tópicos. Essas categorias abrangem uma ampla variedade de áreas, como esportes, política, religião, tecnologia, ciência, entretenimento e muito mais. Cada grupo de notícias contém mensagens de discussão coletadas de fóruns online da Usenet.

- alt.atheism
- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- misc.forsale
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics'
- sci.med
- sci.space
- soc.religion.christian
- talk.politics.guns
- talk.politics.mideast
- talk.politics.misc
- talk.religion.misc