



Problem Statement: Determining how many of the questions in the VQA dataset[1][2] are about common objects (e.g., bike, table, chair, person, toilet, etc.)?

VQA dataset which can be downloaded from VQA challenge webpage[1] consists of two datasets:

1.) **Real Images:** It has 123,287 training and validation images and 81,434 test images from the newly-released Microsoft Common Objects in Context (MS COCO)[10].

2.) **Abstract Scenes:** It consists of 50k scene of cliparts containing 20 “paperdolls”, over 100 objects and 31 animals in various poses[6][7][8][9].

Since for the task to categorizing questions, I am will only discuss about the questions here and not other details like the answers and captions. Albeit, interested readers can find the more information regarding VQA dataset split and protocols adopted for framing questions and answers on [2]. For every image we have 3 questions, in two modalities: open-ended and multiple choice.

Real Image		Abstract Scene	
			
What kind of store is this?	bakery bakery pastry	art supplies grocery grocery	What is the dog looking at? ball ball soccerball
Is the display case as full as it could be?	no no no	no yes yes	Will the boy play with the dog? yes yes yes

Example of Images and corresponding questions in two modalities.

Image Credit: S. Antol et. al. ICCV 2015.

Recipe:

1. First, download the training, validation and testing questions from the download page of [1], which are json files.
2. After extracting the questions, tokenize the sentence using nltk tokenizer[3].
3. Now use nltk POS tagger[3], and filter out all those words which are not proper noun and common noun. Since they are more likely to indicate which categories the question belong to.
4. After this use WordNet lemmatizer [4] for converting the plurals to their singular forms. Still some of plurals will not be converted like 'men' will not be changed 'man' and so on. Hence we can use WordNet morphy [4] for this task.

- Furthermore for representation you can make word clouds from <http://worditout.com/>

[illegible][illegible]

[1] <http://www.visualqa.org>

- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, “VQA: Visual Question Answering”, ICCV 2015. [arXiv:1505.00468v4](https://arxiv.org/abs/1505.00468v4) [cs.CL]
- [3] <http://nltk.org/>

- [4] <http://wordnet.princeton.edu/>
- [5] <http://scikit-learn.org/>
- [6] S. Antol, C. L. Zitnick, and D. Parikh. Zero-Shot Learning via Visual Abstraction. In ECCV, 2014
- [7] C. L. Zitnick and D. Parikh. Bringing Semantics Into Focus Using Visual Abstraction. In CVPR, 2013.
- [8] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the Visual Interpretation of Sentences. In ICCV, 2013.
- [9] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting Abstract Images for Semantic Scene Understanding. PAMI, 2015.
- [10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, 2014.