We collected the cancer data from different hospitals across United States of America. The data is in the form of text format. We want to convert text into numerical data (Text Classification) depending upon the Number of polyps,shape, size, location and treatment method. Each record gives the patient cancer information sometimes the patient may have cancer in multiple locations, so we are trying to have data sorted according to location as parameter.

Data Dictionary

|  |  |
| --- | --- |
| Random \_ ID | Patient ID |
| Number of sessile | Data type: Number(Sessile,

|  |
| --- |
|  |
| PEDUNCULATED |
| FLAT |
| MASS) |

 |
| Size of sessile  | DiminutiveSmall MediumLarge |
| Size of sessile in number  | Number units(MM,CM) |
| Shape of sessile |

|  |
| --- |
| SESSILE |
| PEDUNCULATED |
| FLAT |
| MASS |
| SERRATED |
| SMOOTH |

 |
| Location of sessile  |

|  |  |
| --- | --- |
| Names of location | Other names of location according to data |
| Cecum | Cecal |
| Ascending Colon | Ascending |
| Ileocecal Valve | Ileum/ Ileocecal |
| Hepatic Flexure | Hepatic |
| Transverse Colon | Transverse |
| Splenic Flexure | Splenic |
| Descending Colon | Descending |
| Sigmoid Colon | Sigmoid |
| Recto sigmoid | Recto-sigmoid |
| Rectum | Rectal |
| Appendix | Appendix |

 |
| Side----- location divided into two side left and right | Right Left |
| Treatment  | PiermealCold SnareHot SnareSnare CauteryElectocautery snareSnareExcisional biopsyBiopsy forcepCold biopsyResectionRemovedNot removed --------RetrievalNot Retrieval |

I call this as capsule which contains all the information about the cancer of a specific location. If record says about multiple locations that many capsule should be created. As per my knowledge the highest number locations in one record is 12.



**Instructions**

* Special characters can be removed

Example--, “ \ / ? ! ~ @ # $ % ^ & \* ( ) except>< . –because this signs are mostly used for stating the size of sessile ( 3.5 cm , 3-6 cm ,> 3 cm ,< 4 cm )