Spam Classifier Tutorial

CS3300 Data Science

RJ Nowling

Problem Definition

- We want to predict whether a given email is spam or not.

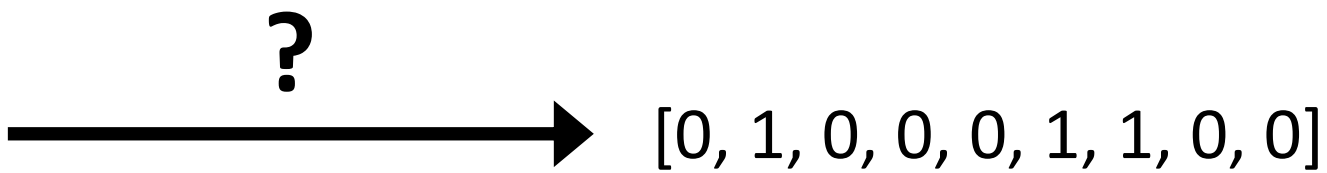
Data Set

- trec07p – University of Waterloo
- ~ 75k emails from between April and July 2007
  - ~25k ham
  - ~50k spam
- Used 75% for training, 25% for testing

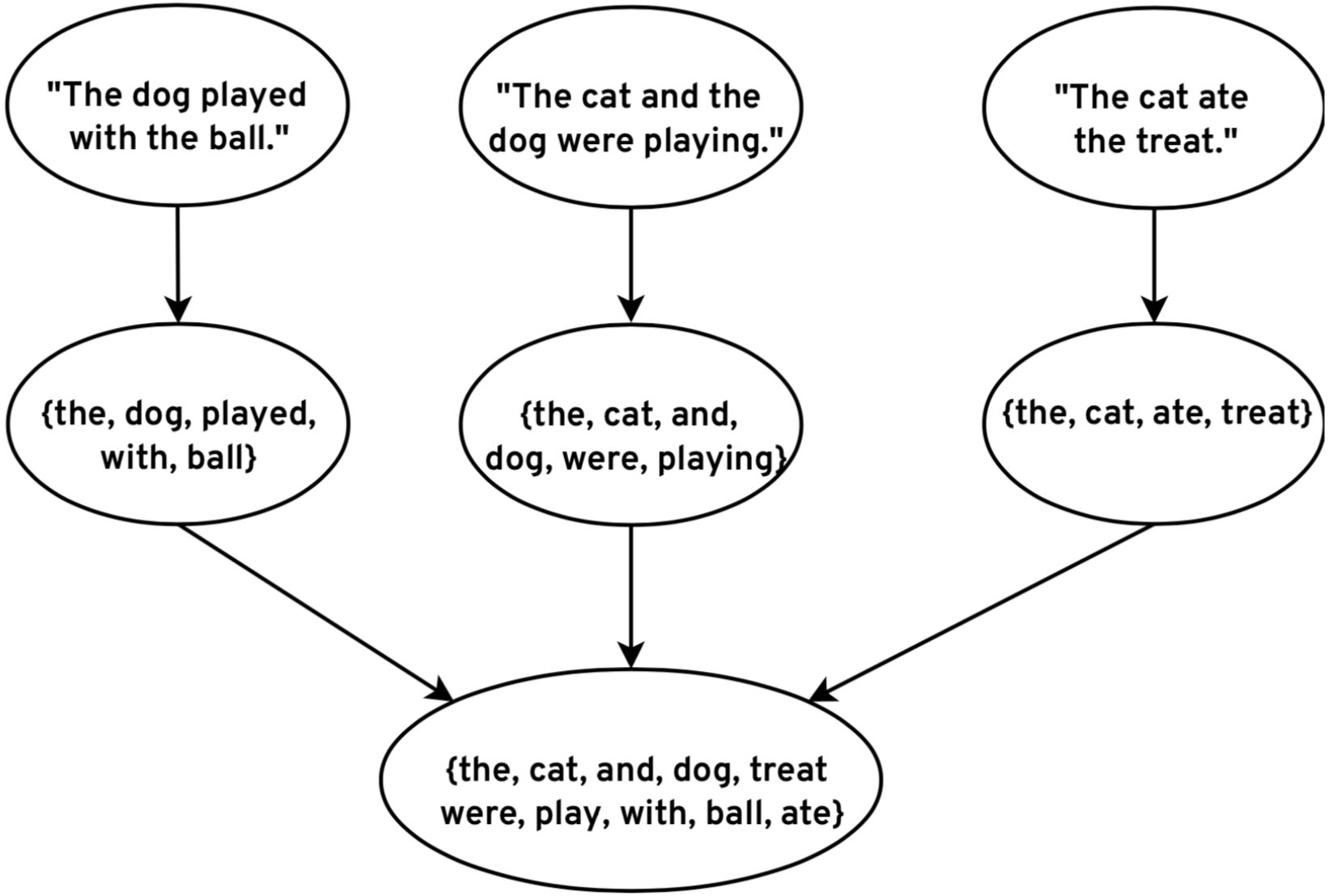
Dear Class,

I've uploaded the new homework to D2L. It's due on Monday. I'm looking forward to seeing your solutions!

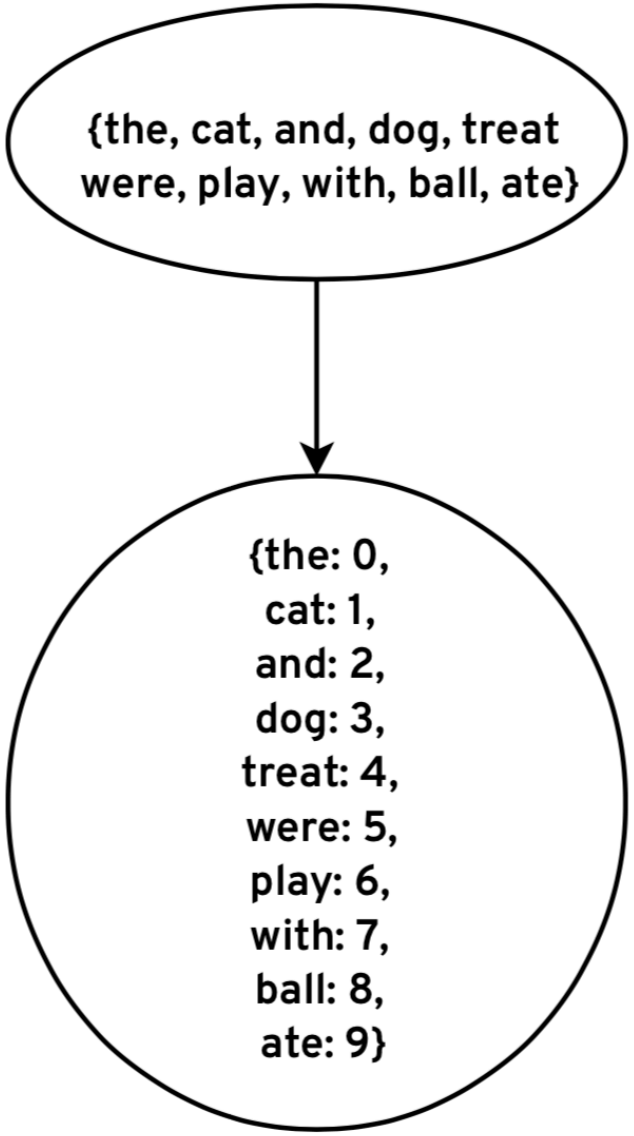
RJ



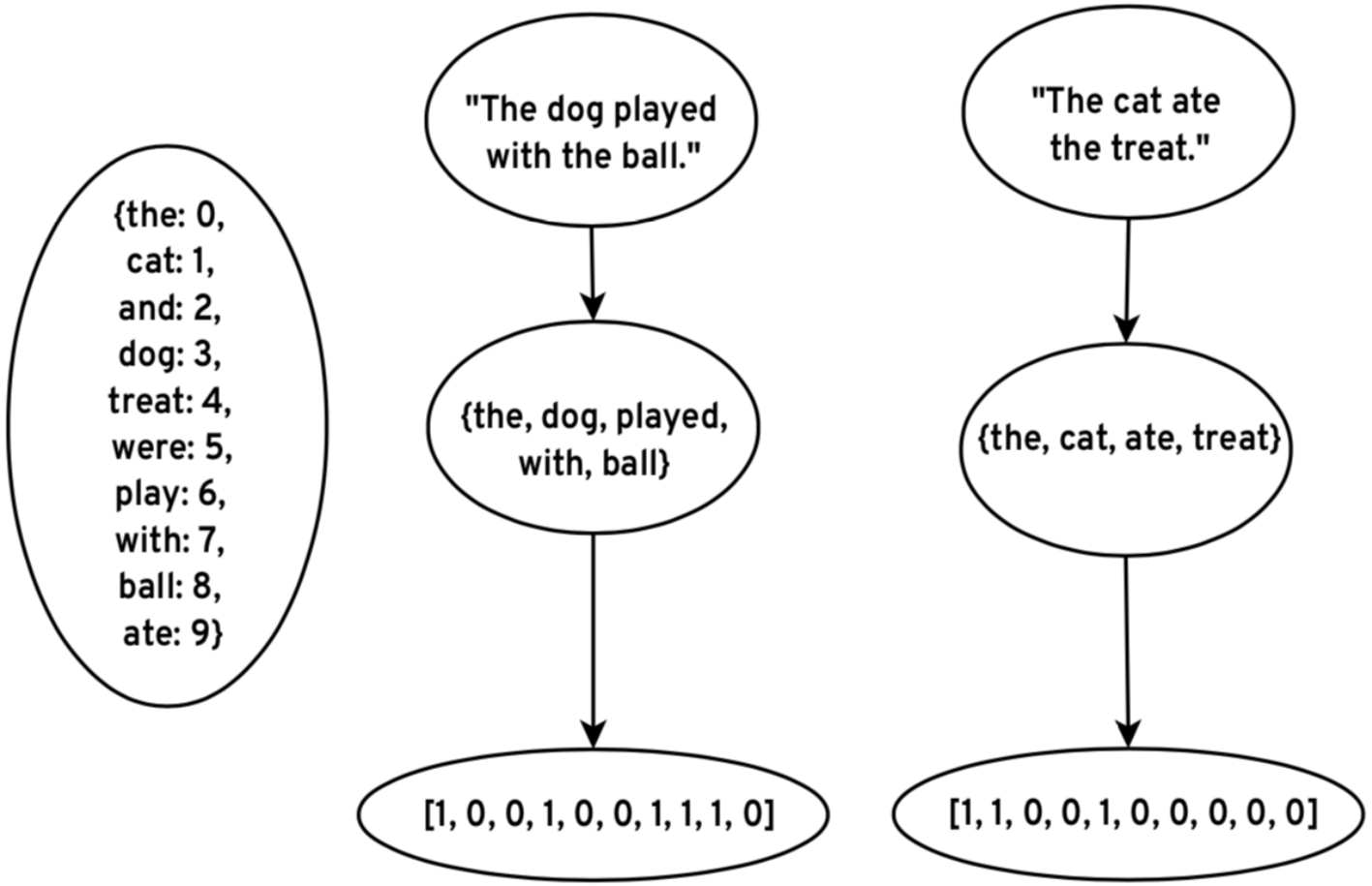
Extract Vocabulary



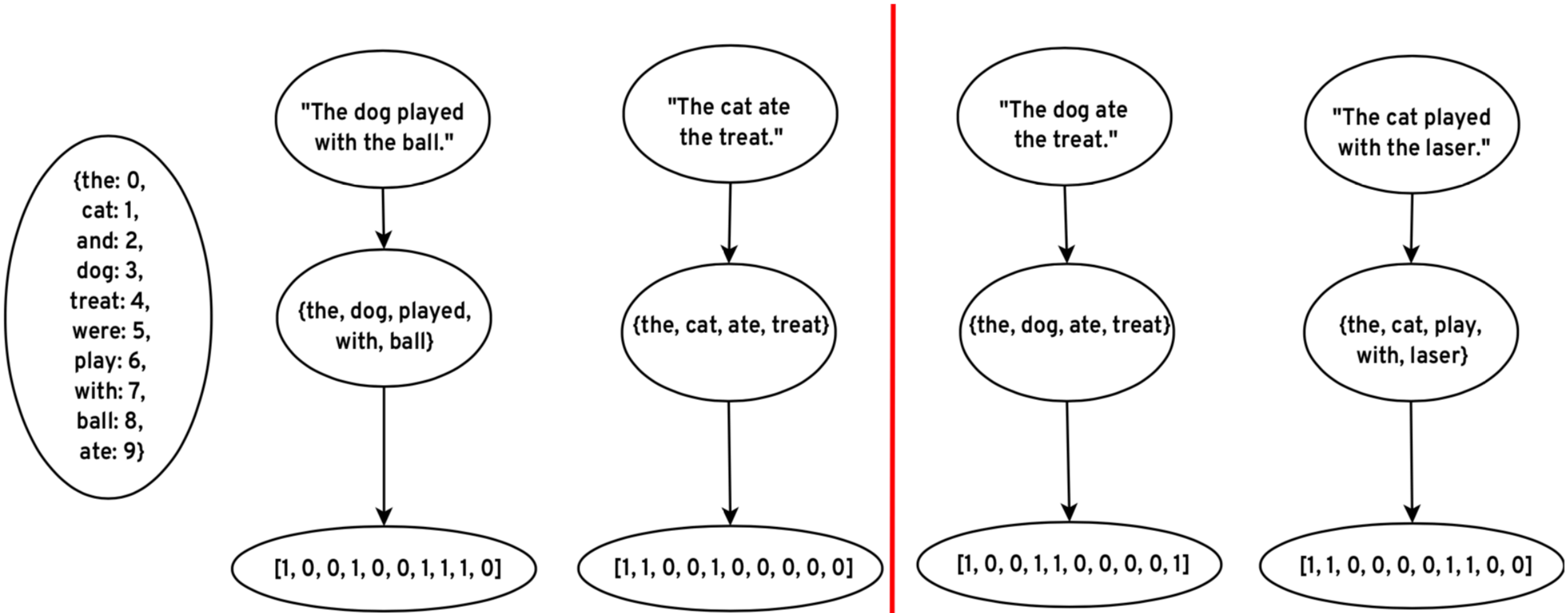
Map Words to Column Indices



Encode Features



Encode Features



```
From xlvwscs@net.il  Wed Apr 11 21:18:50 2007
Return-Path: xlvwscs@net.il
Received: from DSL217-132-183-148.bb.netvision.net.il (89-139-22-
235.bb.netvision.net.il [89.139.22.235])
    by speedy.uwaterloo.ca (8.12.8/8.12.5) with ESMTP id l3C1Im0I024199
    for <gnitpick@speedy.uwaterloo.ca>; Wed, 11 Apr 2007 21:18:49 -0400
From: "repairs" xlvwscs@net.il
To: gnitpick@speedy.uwaterloo.ca
Subject: Secure Web-Form
Date:     Thu, 12 Apr 2007 04:18:33 -0300
MIME-Version: 1.0
Content-Type: multipart/related;
    boundary="----=_NextPart_000_0004_01C77CB9.A20DDD00"
X-Mailer: Microsoft Office Outlook, Build 11.0.5510
Thread-Index: Acd8uaINCus5OCPWRZ2d2pVdMAveNQ==
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.2900.2869
Message-Id: <FC6B2A13C68B036.F7152FB58C@net.il>
```

From: "repairs" xlvwscs@net.il

X-Mailer: Microsoft Office Outlook, Build 11.0.5510

Domains

```
net.il => { net.il, il }
amazon.co.uk => { amazon.co.uk, co.uk, uk }
```

User Agents

```
Microsoft Office Outlook, Build 11.0.5510 =>

{ microsoft, office, outlook, build, 11, 0, 5510 }

Thunderbird 1.5 (Windows/20051201) =>

{ thunderbird, 1, 5, windows, 20051201 }
```