Ling 555 — Programming for Linguists

Version control, edit distance and nltk

Robert Albert Felty

Speech Research Laboratory
Indiana University

Nov. 10, 2008

Outline

1. Homework questions and comments
2. Version Control
   - intro
   - example
   - concepts
   - Two types of version control
   - subversion
3. Edit Distance
   - theory
   - Edit distance usage
4. Natural language toolkit
   - NLTK intro
   - NLTK demo

Version Control intro

Definition

Version control is an essential tool for programmers, providing several key functions:

1. The ability to track code changes
2. The ability to collaborate easily
3. The ability to create and potentially merge different versions of the same project

Also referred to as

- RCS (revision control system)
- SCM (source control management)

A small example

Suppose *Joe the programmer* and I are working on developing a python module.

Joe's copy

```
""" this module does X"""
import re,sys,os,time
foo = 1
bar = 2
```

My copy

```
""" this module does X"""
import re,sys,os
foo = 1
bar = 2
another = 23
```

Basic concepts

revision Every time someone commits something new to the repository, a new revision is created, which is like a snapshot of the project at one particular point in time

repository The repository contains all of the project's files, and most importantly a history of all the changes to it

working copy A working copy is your own personal copy of the repository. It contains only 1 revision of the repository

Version Control types

Centralized

All the code is stored on a central server. Whenever developers want to download the newest version, or upload some changes, they must use the server

- RCS
- CVS (concurrent version system)
- Subversion

Distributed

Every person gets a complete copy of the code, including all the history and changes

- git
- mercurial
- bazaar

subversion intro

Download from subversion.tigris.org

Why subversion?

- Subversion is designed as a replacement for CVS.
- CVS was the most widely-used version control system.
- Subversion is becoming the most widely-used, and fixes lots of problems with CVS.
- free
- available for almost every operating system
- well documented
- relatively easy

subversion commands

`svn help`	Get help on using subversion.
`svn checkout`	Download a fresh copy of a repository
`svn update`	Get the latest updates for your working copy
`svn commit`	Commit some changes you have made to the repository
`svn add`	Add a file or directory to svn (the next time you commit)
`svn mv`	Change the location of a file
`svn diff`	Compare your working copy to the version in the repository

L555 repository

I have created a subversion repository for the class.

```
svn checkout svn://robfelty.com\
/home/robfelty/svn/l555 myl555
```

- There is a subdirectory for each student
- You have read-only permissions on everything
- You have read-write permissions on your own directory

Edit distance

Definition

Edit distance (also known as Levenshtein distance) is the minimal number of additions, deletions, and/or substitutions to change one string into another

Edit distance usage

- DNA sequencing
- plagiarism detection
- measuring phonological similarity
- spell checking
- speech recognition

NLTK intro

- Extensive toolkit for doing / learning computational linguistics
- Written in python
- Includes many corpora
- Has a variety of tools for NLP, tagging, making trees, and grammars
- open-source
- Extensible
- Well-documented
- Actively maintained

NLTK demo

For some nltk demos, look on the delicious page delicious.com/robfelty/l555