تشخیص دزدی ادبی (Plagiarism detection)

دزدی ادبی گونه های مختلفی دارد. در اینجا منظور از دزدی ادبی کپی کردن بخشی از متن دیگران سپس تغییر آن است.

دو نوع متن در اختیار شما است. یک متن اصلی و متن دوم متنی است که از همان متن قبلی کپی شده است. می خواهیم برامه ای بنویسیم که با استفاده از Edit distance تشخیص دهد که آیا یک متن ورودی (تست) کپی شده از یک متن مرجع است یا خیر.

برای این کار، یک متن مرجع و پنج متن ورودی در اختیار شما است. دو تا از این متون (original1,original2) کپی نیستند. و دوتا از آنها کپی شده از متن مرجع هستند. و یکی از متون ورودی (t2) نیز موضوعی کاملا متفاوت با سایرین دارد. درجه دزدی ادبی در یکی از متون کپی شده (copy) بسیار بالا و در دیگری (revised) پایین است. در اولی، متن عینا کپی شده و فقط برخی از جملات حذف شده یا اندکی تغییر کرده اند ولی در دومی تغییرات اعمال شده بسیار وسیع بوده حتی جملات و پاراگراف ها جابه جا شده اند.

برنامه شما باید برای هریک از پنج متن ورودی یک عدد را محاسبه نماید که درجه دزدی ادبی را درآنها نسبت به متن مرجع، نشان دهد. سپس یک حد آستانه را برای تشخیص اینکه دزدی ادبی اتفاق افتاده است یا خیر تعیین نمایید.

انتساب عدد مذکور به هر متن باید از Edit distance استفاده نمایید ولی می توانید در کنار آن از معیارهای دیگری نیز بهره ببرید. به عنوان مثال می توانید تعداد کلمات مشترک را در دو متن بشمارید. استفاده درست از معیارهای اضافی منجر به دریافت نمره اضافه خواهد شد.